

**МЕТОДЫ  
СТАТИСТИЧЕСКОЙ ОБРАБОТКИ  
ИНФОРМАЦИИ В MS EXCEL**

**Пособие  
для студентов специальности 1-26 03 01  
«Управление информационными ресурсами»**

УДК 004  
ББК 32.973.202  
М 54

Авторы-составители: О. И. Еськова, канд. техн. наук, доцент;  
Л. П. Авдашкова, канд. физ.-мат. наук, доцент

Рецензенты: Л. Н. Марченко, канд. техн. наук, доцент  
Гомельского государственного университета  
им. Ф. Скорины;  
Е. А. Левчук, канд. техн. наук, доцент  
Белорусского торгово-экономического  
университета потребительской кооперации

Рекомендовано к изданию научно-методическим советом учреждения образования «Белорусский торгово-экономический университет потребительской кооперации». Протокол № 2 от 13 декабря 2011 г.

Методы статистической обработки информации в MS Excel : пособие для студентов специальности 1-26 03 01 «Управление информационными ресурсами» / авт.-сост.: О. И. Еськова, Л. П. Авдашкова. – Гомель : учреждение образования «Белорусский торгово-экономический университет потребительской кооперации», 2012. – 132 с.  
ISBN 978-985-461-938-5

УДК 004  
ББК 32.973.202

ISBN 978-985-461-938-5

© Учреждение образования «Белорусский  
торгово-экономический университет  
потребительской кооперации», 2012

Одной из особенностей экономических систем является значительное влияние случайных факторов на исследуемые показатели и, следовательно, необходимость учета этого влияния при принятии управленческих решений. Поэтому методы статистического анализа экономической информации изучаются в рамках дисциплины «Экономико-математические методы и модели принятия решений», предназначенной для студентов специальности 1-26 03 01 «Управление информационными ресурсами», целью которой является систематизация знаний по математической статистике, полученных при изучении высшей математики, а также практическое применение этих знаний для задач прикладного характера. Поэтому основная направленность данного пособия – продемонстрировать возможности для статистической обработки табличного процессора MS Excel, который в настоящее время является одним из наиболее популярных средств анализа данных экономических наблюдений. Не умаляя важности специализированных программ для обработки статистических данных (например, пакета Statistica), авторы считают, что изучение интерфейса и возможностей таких объемных и мощных программных продуктов в рамках одного раздела дисциплины займет слишком много времени и отвлечет внимание студентов от сути статистических методов, условий и способов их применения. Поэтому был сделан выбор в пользу MS Excel как уже хорошо знакомого студентам инструмента, возможности которого для статистической обработки данных также весьма велики.

Настоящее издание включает три главы.

- *Первая глава* «Случайные величины» посвящена базовым понятиям теории вероятностей, где рассматриваются понятия случайного события, вероятности и случайной величины, способы описания законов распределения случайных величин, расчета их основных характеристик. Кроме того, здесь описываются некоторые, наиболее часто встречающиеся, законы распределения случайных величин. При хорошей математической подготовке студентов данная глава может быть пропущена либо использована для повторения отстающими студентами.

- Во *второй главе* «Выборка и ее анализ» представлены основные подходы к анализу выборочных данных: построение вариационного ряда, полигона частот и гистограммы, расчет выборочных характеристик и анализ полученных значений, проверка гипотезы о соответствии выборочных данных некоторому известному закону распределения. Все этапы анализа выборки могут быть автоматизированы средствами MS Excel. В этой главе излагаются также основные понятия и принципы проверки статистических гипотез.

- *Третья глава* «Анализ нескольких выборок» позволяет сформировать у студентов представление о сравнительном анализе нескольких случайных величин на основе выборочных данных. Здесь рассматриваются такие направления анализа нескольких выборок, как корреляционный и регрессионный анализ, способы сравнения математических ожиданий и дисперсий двух выборок, а также оценка влияния некоторого фактора на случайную величину с помощью дисперсионного анализа. Все рассматриваемые методы сравнения выборок могут быть автоматизированы с помощью инструментов надстройки *Пакет анализа*, причем большое внимание уделяется особенностям использования этих инструментов и их правильному выбору.

В пособие включено большое количество примеров реализации описываемых методов в MS Excel. Даны задания для самостоятельной работы, что позволяет использовать данное издание в качестве практикума для лабораторных работ. При этом теоретический материал полностью соответствует лекционному курсу по разделу дисциплины «Экономико-математические методы и модели принятия решений» и может использоваться для самостоятельной работы студентов.

Пособие может быть также использовано магистрантами, аспирантами и студентами других специальностей, если характер их научной работы предполагает обработку результатов экономических наблюдений.

**ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ,  
МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО РЕШЕНИЮ ПРИМЕРОВ,  
ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ**

## **1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ**

### **1.1. Случайные события и вероятность**

В теории вероятностей соблюдение совокупности условий при проведении эксперимента называется *испытанием*. *Событие* является результатом (или исходом) испытания.

События можно разделить на три вида: достоверные, невозможные и случайные.

*Достоверным* называется событие, которое обязательно произойдет, если будет осуществлена определенная совокупность условий. Например, если подброшен игральный кубик, то событие «кубик упадет вниз» есть достоверное.

Событие называется *невозможным*, если оно никогда не произойдет при совокупности данных условий. При подбрасывании кубика событие «кубик улетит в космос» является невозможным.

Событие называется *случайным*, если в результате наблюдения или испытания оно может произойти или не произойти. Например, «выпадение одного очка», «выпадение двух очков» и т. д. при подбрасывании кубика – случайные события.

В примере с кубиком также предполагалось, что выполняется определенная совокупность условий: подбрасывание происходит на Земле, а не в невесомости, кубик изготовлен из однородного материала, центр его тяжести не смещен и т. д.

**Пример 1.1.** В урне имеются цветные шары. Из урны наудачу берут один шар. Извлечение шара из урны есть испытание. Появление шара определенного цвета – событие. Хотя обычно это не оговаривается, но предполагается, что здесь соблюдаются определенные условия: шары тщательно перемешаны, шар берут «не глядя» (с закрытыми глазами) и т. п.

События называются *несовместными*, если появление одного из них исключает появление других событий в одном и том же испытании. Например, «выпадение двух очков» и «выпадение трех очков» при подбрасывании кубика являются несовместными событиями.

Несколько событий образуют *полную группу*, если в результате испытания должно появиться хотя бы одно из них. При этом, если события, образующие полную группу, попарно несовместны, то в результате испытания появится одно и только одно из этих событий.

*Противоположными* называются два единственно возможных события, образующих полную группу. Если одно из двух противоположных событий обозначено через  $A$ , то другое принято обозначать  $\bar{A}$ .

**Пример 1.2.** При подбрасывании кубика шесть событий («выпадение одного очка», «выпадение двух очков» ... «выпадение шести очков») попарно несовместны и образуют полную группу.

События «выпадение четного числа очков» и «выпадение нечетного числа очков» также образуют полную группу. Более того, они являются противоположными.

**Пример 1.3.** Из ящика наудачу взята деталь. События «взята деталь стандартная» и «взята деталь нестандартная» – противоположные события.

События называются *равновозможными*, если есть основание считать, что ни одно из них не является более возможным, чем другое.

В примере с бросанием кубика события «выпадение одного очка», «выпадение двух очков» и т. д. являются равновозможными, если речь идет об обычном кубике с несмещенным центром тяжести. Но если, например, прилепить к одной из граней этого кубика кусочек жевательной резинки, то эти события перестанут быть равновозможными.

Основной характеристикой случайного события является его вероятность. *Вероятностью события  $A$*  называется число, характеризующее степень возможности появления события в испытании. Существует несколько способов определения вероятности. Следует рассмотреть подход, который называется классическим.

Каждый из возможных результатов испытания назовем *элементарным событием* (*элементарным исходом*). Те элементарные события, при которых интересующее нас событие наступает, назовем *благоприятствующими* этому событию.

Вероятность события  $A$  рассчитывается как отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу. Таким образом, вероятность события  $A$  определяется формулой

$$P(A) = m/n, \quad (1.1)$$

где  $m$  – число элементарных исходов, благоприятствующих  $A$ ;

$n$  – число всех возможных элементарных исходов испытания.

**Пример 1.4.** Пусть нас интересует событие «выпадение четного числа очков» при подбрасывании кубика (событие  $A$ ). Всего имеется шесть элементарных исходов данного испытания ( $n = 6$ ), так как кубик может упасть на любую из шести своих граней. Благоприятствующие событию  $A$  будут следующие элементарные исходы: «выпадение двух очков», «выпадение четырех очков» и «выпадение шести очков», т. е.  $m = 3$ . Таким образом,  $P(A) = 3/6 = 1/2$ .

**Пример 1.5.** Пусть в урне содержится 6 одинаковых, тщательно перемешанных шаров, причем 2 из них – красные, 3 – синие и 1 – белый. Наудачу вытаскивают из урны один шар. Рассмотрим вероятность события  $A$  = «появился красный шар».

*Решение*

Поскольку возможно вытащить любой из шести шаров, то элементарных исходов в одном испытании ( $n$ ) равно 6. Благоприятствующих рассматриваемому событию исходов ( $m$ ) будет равно 2, поскольку в урне два красных шара и появление любого из них означает наступление события  $A$ . Поэтому вероятность события  $A$  равна:  $P(A) = 2/6$ .

Из определения вероятности вытекают следующие ее свойства:

1. Вероятность достоверного события равна 1.
2. Вероятность невозможного события равна 0.
3. Вероятность случайного события есть положительное число, заключенное между нулем и единицей.

Суммой  $A + B$  событий  $A$  и  $B$  называют событие, состоящее в появлении события  $A$  или события  $B$ , или обоих этих событий.

Например, если из орудия произведены два выстрела и  $A$  – попадание при первом выстреле,  $B$  – попадание при втором выстреле, то  $A + B$  – попадание при первом выстреле или при втором выстреле, или при обоих выстрелах.

**Теорема 1.** Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий:

$$P(A+B) = P(A) + P(B).$$

**Теорема 2.** Сумма вероятностей несовместных событий  $A_1, A_2, \dots, A_n$ , образующих полную группу, равна единице:

$$P(A_1) + P(A_2) + \dots + P(A_n) = 1.$$

**Теорема 3.** Сумма вероятностей противоположных событий равна единице:

$$P(A) + P(\bar{A}) = 1.$$

Если вероятность одного из двух противоположных событий обозначена через  $p$ , то вероятность другого события обозначается через  $q$ . Таким образом,  $p + q = 1$ .

**Пример 1.6.** Пусть вероятность того, что взятая из ящика деталь является стандартной ( $p$ ), равна 0,9. Найти вероятность того, что деталь будет нестандартной ( $q$ ).

*Решение*

События «появилась стандартная деталь» и «появилась нестандартная деталь» являются противоположными, поэтому искомая вероятность будет равна:

$$q = 1 - p = 1 - 0,9 = 0,1.$$

Произведением двух событий  $A$  и  $B$  называют событие  $AB$ , состоящее в их совместном появлении. Например, если  $A$  – «деталь годная», а  $B$  – «деталь окрашенная», то  $AB$  – «деталь годная и окрашенная».

Если при вычислении вероятности события никаких других ограничений, кроме необходимого комплекса условий испытания, не налагается, то такая вероятность называется безусловной. Если же налагаются другие дополнительные условия, то вероятность события называется условной.

Условной вероятностью  $P_A(B)$  называют вероятность события  $B$ , вычисленную в предположении, что событие  $A$  уже наступило.

**Пример 1.7.** Пусть в урне (см. пример 1.5) содержатся 2 красных, 3 синих и 1 белый шар. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления синего шара при втором испытании (событие  $B$ ) при условии, что при первом испытании извлечен красный шар (событие  $A$ ).

*Решение*

После первого извлечения в урне осталось 5 шаров, из них – 3 синих. Искомая условная вероятность следующая:  $P_A(B) = 3/5$ .

**Теорема 4.** Вероятность произведения двух событий определяется формулой

$$P(AB) = P(A)P_A(B).$$

**Пример 1.8.** Используя условие примера 1.7, найти вероятность того, что первый шар будет красный, а второй – синий.

*Решение*

В примере 1.5 была найдена вероятность того, что при первом извлечении появится красный шар:  $P(A) = 2/6$ . В примере 1.7 была найдена условная вероятность вытянуть синий шар после красного:  $P_A(B) = 3/5$ . Поэтому искомая вероятность произведения событий будет равна:

$$P(AB) = P(A)P_A(B) = \frac{2}{6} \cdot \frac{3}{5} = \frac{2}{10}.$$

### **Задания для самостоятельной работы**

1. Консультационный пункт института получает пакеты с контрольными работами из городов А, В и С. Вероятность получения пакета из города А равна 0,7, из города В – 0,2. Найдите вероятность того, что очередной пакет будет получен из города С.
2. Набирая номер телефона, абонент забыл одну цифру и набрал ее наудачу. Найдите вероятность того, что набрана нужная цифра.
3. В ящике имеется 50 одинаковых деталей, из них – 5 окрашенных. Наудачу вынимают одну деталь. Найдите вероятность того, что извлеченная деталь окажется окрашенной.
4. Участники жеребьевки тянут из ящика жетоны с номерами от 1 до 100. Найдите вероятность того, что номер первого наудачу извлеченного жетона не содержит цифры 5.
5. На основании условия примера 1.5 найдите вероятность того, что шар окажется цветным (красным или синим).
6. Стрелок стреляет по мишени, разделенной на 3 области. Вероятность попадания в первую область равна 0,45, во вторую – 0,35. Найдите вероятность того, что стрелок при одном выстреле попадет либо в первую, либо во вторую область.
7. В денежно-вещевой лотерее на каждые 10 000 билетов разыгрывается 150 вещевых и 50 денежных выигрышей. Определите, чему равна вероятность выигрыша, безразлично денежного или вещевого, для владельца одного лотерейного билета.
8. В ящике находится 10 деталей, из них – 2 нестандартных. Из ящика наудачу извлекают одну деталь, а затем вторую. Найдите вероятность того, что обе детали окажутся стандартными.

## **1.2. Способы описания и характеристики случайных величин**

Если со случайным событием связать некоторое числовое значение, то приходим к понятию случайной величины. Например, при бросании игральной кости число выпавших очков будет случайной величиной, которая может принимать одно из значений от 1 до 6.

*Случайной величиной (СВ)* называется величина  $X$ , которая в результате испытания может принимать то или иное значение, причем заранее (до испытания) не известное.

Каждой случайной величине соответствует множество чисел – это множество значений, которые она может принимать. Будем обозначать случайные величины прописными буквами  $X, Y, Z$  и т. д., а значения этих величин – строчными буквами  $x, y, z$  и т. д.

Различают два вида случайных величин: дискретные и непрерывные.

*Дискретной* называют случайную величину, которая принимает отдельные, изолированные возможные значения с определенными вероятностями. Число возможных значений дискретной случайной величины может быть конечным и бесконечным. Например, число родившихся мальчиков из ста новорожденных есть дискретная случайная величина, которая принимает конечное множество значений: 0, 1, 2 ... 100. Число людей, зашедших в дверь магазина за один час, также является дискретной случайной величиной и принимает значения 0, 1, 2 и т. д. При этом нельзя назвать предельное количество людей, которые могут зайти, т. е. множество значений является бесконечным.

*Непрерывной* называется случайная величина, которая может принимать все значения из некоторого числового промежутка. Очевидно, что число возможных значений непрерывной случайной величины бесконечно. Например, некоторая деталь, производимая на токарном станке, имеет длину 7 см. По ГОСТу допустимы отклонения этой длины в обе стороны, не превышающие 2 мм. Таким образом, длина стандартной детали является случайной величиной и может принимать любое значение из интервала  $[6,8;7,2]$  (сантиметров).

*Законом распределения дискретной случайной величины* называется соответствие между отдельными возможными ее значениями и их вероятностями. Этот закон может быть задан в виде таблицы, формулы или графика. В случае табличного задания закона распределения в первой строке таблицы указываются возможные значения случайной величины, а во второй – их вероятности (таблица 1.1).

Таблица 1.1 – Закон распределения дискретной случайной величины

$X$	$x_1$	$x_2$	$\dots$	$x_n$
$P$	$p_1$	$p_2$	$\dots$	$p_n$

Поскольку в одном испытании случайная величина принимает только одно возможное значение, то события  $X = x_1, X = x_2, \dots, X = x_n$  образуют полную группу, т. е. сумма их вероятностей равна единице:

$$\sum_{i=1}^n p_i = 1. \quad (1.2)$$

**Пример 1.9.** В денежной лотерее на 100 билетов разыгрывается один выигрыш стоимостью 20 р., два выигрыша – по 10 р. и 10 выигрышей – по 1 р. Найти закон распределения случайной величины  $X$  – возможного выигрыша на один билет.

*Решение*

Возможные значения выигрышей составляют:  $x_1 = 0, x_2 = 1, x_3 = 10$  и  $x_4 = 20$ . Для расчета вероятности каждого выигрыша разделим число благоприятствующих соответствующему выигрышу исходов на число всех исходов:  $p_4 = 1/100 = 0,01; p_3 = 2/100 = 0,02; p_2 = 10/100 = 0,1$ . Очевидно, что число билетов с нулевым выигрышем будет следующим:  $100 - 1 - 2 - 10 = 87$ ; поэтому  $p_1 = 87/100 = 0,87$ . Проведем контроль:  $0,87 + 0,1 + 0,02 + 0,01 = 1$ . Искомый закон распределения имеет вид, показанный в таблице 1.2.

Таблица 1.2 – Закон распределения выигрыша на один билет

$X$	0	1	10	20
$P$	0,87	0,1	0,02	0,01

В случае графического задания дискретной случайной величины по оси  $OX$  откладывают значения этой величины, а по оси  $OY$  – соответствующие вероятности. Ломаная линия, соединяющая точки  $(x_i, p_i)$ , называется *многоугольником распределения* (рисунок 1.1).

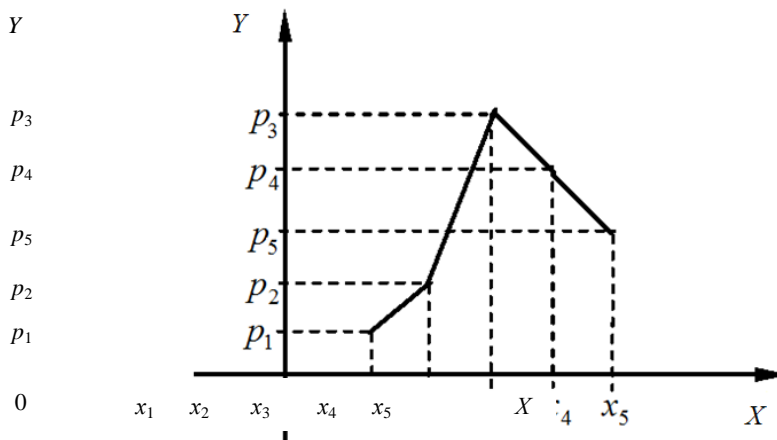


Рисунок 1.1 – Многоугольник распределения дискретной случайной величины

Закон распределения дискретной случайной величины может быть задан также в виде формулы, отражающей зависимость вероятности от значения случайной величины:

$$p_i = P(x_i).$$

*Функцией распределения* случайной величины  $X$  называется функция  $F(x)$ , определяющая вероятность того, что случайная величина  $X$  в результате испытания примет значение, меньшее  $x$ :

$$F(x) = P(X < x).$$

Геометрически это означает, что случайная величина  $X$  примет значение, изображаемое на числовой оси точкой, расположенной левее точки  $x$ . Иногда вместо термина «функция распределения» используется термин «интегральная функция».

Функция распределения характеризуется следующими *свойствами*:

1. Функция  $F(x)$  принимает значения от 0 до 1:

$$0 \leq F(x) \leq 1.$$

2. Функция  $F(x)$  – неубывающая функция, т. е.:

$$F(x_2) \geq F(x_1), \text{ если } x_2 > x_1.$$

3. Вероятность того, что случайная величина  $X$  примет значение, заключенное в интервале  $[a, b)$ , равна приращению функции распределения на этом интервале:

$$P(a \leq X < b) = F(b) - F(a).$$

4. Если все возможные значения случайной величины принадлежат интервалу  $[a, b]$ , то

$$F(x) = 0, \text{ если } x \leq a;$$

$$F(x) = 1, \text{ если } x > b.$$

Для дискретной случайной величины функция распределения вычисляется по следующей формуле:

$$F(x) = \sum_{x_i \leq x} p_i. \quad (1.3)$$

**Пример 1.10.** Построить функцию распределения для случайной величины, используя условие примера 1.9 (см. таблицу 1.2).

*Решение*

Если  $x \leq 0$ , то  $F(x) = 0$  (четвертое свойство). Действительно, если взять любую точку из данного интервала, то не существует ни одного значения случайной величины «выигрыш на один билет», которое было бы меньше.

Если  $0 < x \leq 1$ , то  $F(x) = 0,87$ . Зафиксируем любое число  $x_1$  из данного интервала. Меньше него только одно значение случайной величины (равное 0). Поэтому вероятность того, что значение случайной величины будет меньше  $x_1$  – это вероятность того, что оно будет равно 0.

Если  $1 < x \leq 10$ , то  $F(x) = 0,87 + 0,1 = 0,97$ . Действительно, если  $x$  удовлетворяет неравенству  $1 < x \leq 10$ , то событие  $X < x$  может быть осуществлено, когда  $X$  примет значение 0 (вероятность этого события равна 0,87) или значение 1 (вероятность этого события равна 0,1). Поскольку эти два события несовместны, вероятность того, что произойдет одно из них, равно сумме вероятностей ( $0,87 + 0,1 = 0,97$ ).

Если  $10 < x \leq 20$ , то  $F(x) = 0,87 + 0,1 + 0,02 = 0,99$ . Аналогично предыдущему случаю для любой точки  $x$  из данного интервала событие  $X < x$  может наступить в трех случаях:  $X$  примет значение 0 (вероятность этого события равна 0,87),  $X$  примет значение 1 (вероятность равна 0,1) и  $X$  примет значение 10 (вероятность равна 0,02). Поэтому нужно найти сумму этих трех величин.

Если  $x > 20$ , то  $F(x) = 1$ . Действительно, все возможные значения случайной величины «выигрыш на один билет» меньше числа  $x$  из данного интервала, т. е. событие  $X < x$  является достоверным. Следовательно, его вероятность равна единице.

Итак, функция распределения аналитически может быть записана в виде:

$$F(x) = \begin{cases} 0 & \text{при } x \leq 0, \\ 0,87 & \text{при } 0 < x \leq 1, \\ 0,97 & \text{при } 1 < x \leq 10, \\ 0,99 & \text{при } 10 < x \leq 20, \\ 1 & \text{при } x > 20. \end{cases}$$

График функции распределения приведен на рисунке 1.2.



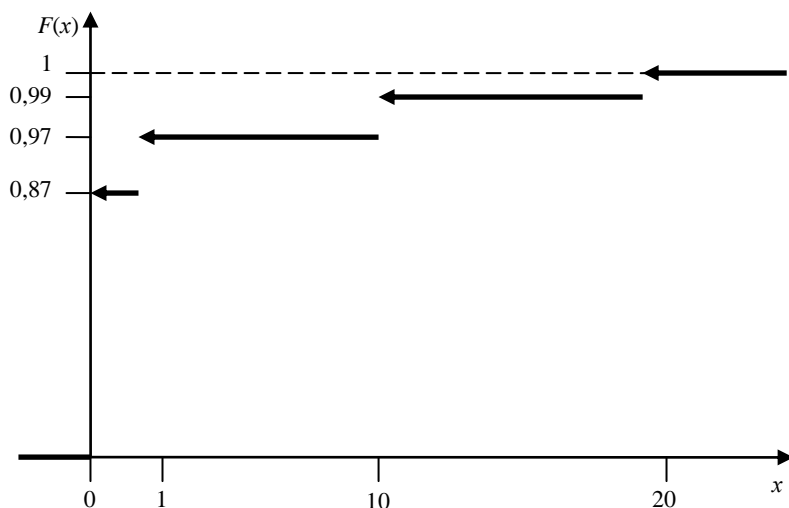


Рисунок 1.2 – График функции распределения дискретной случайной величины

Очевидно, что график функции распределения дискретной случайной величины всегда имеет ступенчатый вид.

С помощью функции распределения можно описать не только дискретную, но и непрерывную случайную величину. Все перечисленные выше свойства функции распределения выполняются и в этом случае. График функции распределения непрерывной случайной величины имеет вид, показанный на рисунке 1.3. Он заключен в полосу, ограниченную прямыми:  $y = 0$  и  $y = 1$  (первое свойство). При возрастании  $x$  в интервале  $[a, b]$ , в котором заключены все возможные значения случайной величины, график «поднимается вверх» (второе свойство). При  $x \leq a$  ординаты графика равны нулю; при  $x > b$  ординаты графика равны единице (четвертое свойство).

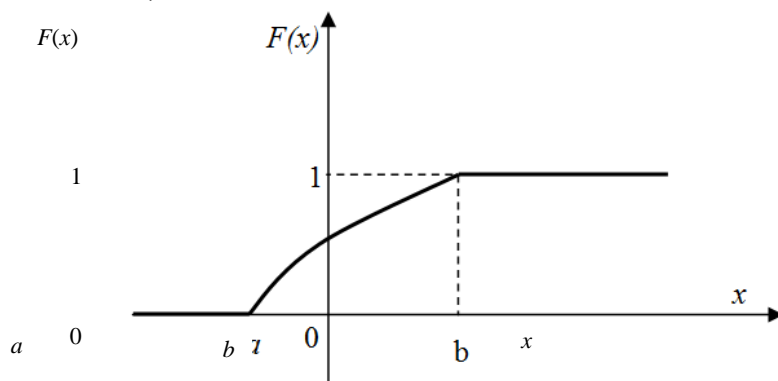


Рисунок 1.3 – Функция распределения непрерывной случайной величины

Непрерывную случайную величину можно также задать с помощью функции плотности распределения (дифференциальной функции распределения).

*Плотностью распределения вероятностей* непрерывной случайной величины  $X$  называется первая производная от функции распределения:

$$f(x) = F'(x).$$

Функция распределения является первообразной для плотности распределения, поэтому

$$P(X < x) = F(x) = \int_{-\infty}^x f(z) dz.$$

Вероятность попадания непрерывной случайной величины в интервал от  $a$  до  $b$  определяется по формуле

$$P(a \leq x \leq b) = F(b) - F(a) = \int_a^b f(x) dx.$$

Геометрически это можно истолковать следующим образом: вероятность того, что случайная величина попадет в интервал от  $a$  до  $b$  равна площади криволинейной трапеции, ограниченной осью  $OX$ , кривой плотности распределения  $f(x)$  и вертикальными прямыми  $x = a$  и  $x = b$  (рисунок 1.4).

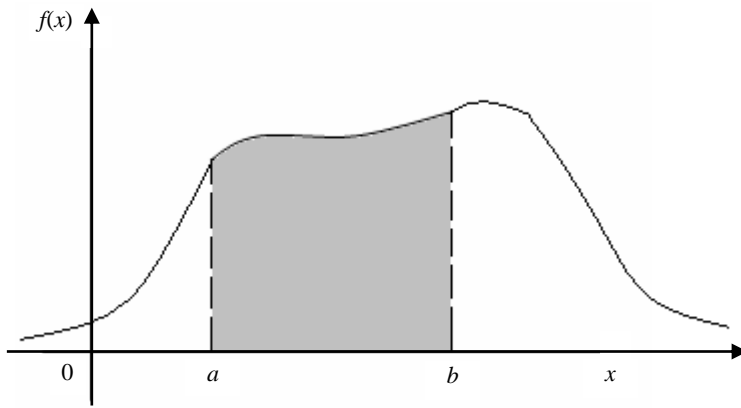


Рисунок 1.4 – Геометрическая иллюстрация вероятности попадания случайной величины в заданный интервал

Функция плотности распределения вероятностей характеризуется следующими *свойствами*:

1. Плотность распределения является неотрицательной величиной:  $f(x) \geq 0$ .

2.  $\int_{-\infty}^{+\infty} f(x)dx = 1$ . Геометрически это означает, что вся площадь криволинейной трапеции, ограниченной

осью  $OX$  и кривой плотности распределения, равна единице.

Таким образом, для описания дискретной случайной величины используется закон распределения и функция распределения. А для описания непрерывной случайной величины используется также функция распределения и функция плотности распределения. Наряду с этим оба типа случайных величин характеризуются с помощью некоторых чисел, которые описывают случайную величину в целом. Такие числа называются числовыми характеристиками случайной величины. К ним относятся математическое ожидание, дисперсия, среднее квадратическое отклонение и др.

*Математическое ожидание* приближенно равно среднему значению случайной величины.

Математическое ожидание может быть рассчитано по следующим формулам:

- для дискретной случайной величины с конечным множеством значений:

$$M(X) = \sum_{i=1}^n x_i p_i;$$

- для дискретной случайной величины с бесконечным множеством значений:

$$M(X) = \sum_{i=1}^{\infty} x_i p_i;$$

- для непрерывной случайной величины:

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

*Дисперсия* характеризует степень рассеяния значений случайной величины вокруг ее математического ожидания. Дисперсия равна математическому ожиданию квадрата отклонения случайной величины от ее математического ожидания:

$$D(X) = M[X - M(X)]^2.$$

Приведенную формулу можно преобразовать к виду, удобному для вычислений:

$$D(X) = M(X^2) - [M(X)]^2. \quad (1.4)$$

*Средним квадратическим отклонением* случайной величины называется квадратный корень из ее дисперсии:

$$\sigma(X) = \sqrt{D(X)}. \quad (1.5)$$

Среднее квадратическое отклонение измеряется в тех же единицах, что и значение случайной величины  $X$ , поэтому с его помощью удобнее характеризовать рассеяние случайной величины относительно его математического ожидания.

*Модой* дискретной случайной величины ( $Mo[X]$ ) называется наиболее вероятное ее значение. Для непрерывной случайной величины мода – это точка максимума ее плотности вероятности.

*Медианой* произвольной случайной величины ( $Me[X]$ ) называется такое ее значение, относительно которого равновероятно получение большего или меньшего значений.

**Пример 1.11.** Найти числовые характеристики случайной величины «выигрыш на один билет» (см. таблицу 1.2).

### Решение

Искомое математическое ожидание равно сумме произведений всех возможных значений случайной величины на их вероятности:

$$M(X) = 0 \cdot 0,87 + 1 \cdot 0,1 + 10 \cdot 0,02 + 20 \cdot 0,01 = 0,5.$$

Это значение характеризует средний выигрыш, приходящийся на один билет. Для расчета дисперсии определяем математическое ожидание величины  $X^2$ :

$$M(X^2) = 0^2 \cdot 0,87 + 1^2 \cdot 0,1 + 10^2 \cdot 0,02 + 20^2 \cdot 0,01 = 6,1.$$

Рассчитываем дисперсию:

$$D(X) = M(X^2) - [M(X)]^2 = 6,1 - 0,5^2 = 5,85.$$

Среднее квадратическое отклонение равно:  $\sigma(X) = \sqrt{5,85} = 2,42$ .

Модой этой случайной величины является значение  $Mo[X] = 0$ .

Медиану для этой задачи вычислить нельзя.

### Задания для самостоятельной работы

1. Из календаря наугад был вырван листок. Постройте закон распределения случайной величины  $X$ , которая обозначает число, указанное на этом листке (считайте, что в году 365 дней). Постройте функцию распределения. Рассчитайте числовые характеристики случайной величины.

2. Возможные значения случайной величины таковы:  $x_1 = 2$ ,  $x_2 = 5$ ,  $x_3 = 8$ . Известны вероятности первых двух возможных значений:  $p_1 = 0,4$ ,  $p_2 = 0,15$ . Найдите вероятность  $x_3$ . Постройте функцию распределения. Рассчитайте числовые характеристики этой случайной величины.

3. В таблице 1.3 представлено распределение годовой прибыли ( $X$ ) фирмы. Определите числовые характеристики этого распределения, постройте функцию распределения и рассчитайте вероятность положительной прибыли.

Таблица 1.3 – Закон распределения годовой прибыли фирмы

$X$	-10	-5	0	10	20	25
$P$	0,05	0,15	0,25	0,30	0,20	0,05

4. Пусть ежедневные расходы на обслуживание и рекламу автомобилей в автосалоне составляют в среднем 100 усл. ед., а число продаж автомашин в течение дня ( $X$ ) подчиняется закону распределения, показанному в таблице 1.4. Цена за машину составляет 5 000 усл. ед., а ее закупочная стоимость равна 4 750 усл. ед. Найдите математическое ожидание и среднее квадратическое отклонение ежедневной прибыли автосалона.

Таблица 1.4 – Закон распределения числа продаж автомашин

$X$	0	1	2	3	4	5	6	7	8	9
$P$	0,25	0,2	0,1	0,1	0,1	0,1	0,05	0,05	0,025	0,025

## 1.3. Основные виды распределений случайных величин

### 1.3.1. Биномиальное распределение

Биномиальное распределение является дискретным законом распределения. Пусть производится  $n$  независимых испытаний, и в каждом из них событие  $A$  может появиться либо не появиться. Вероятность наступления события в каждом испытании постоянна (испытания производятся в одинаковых условиях) и равна  $p$ . Рассмотрим число появления события  $A$  в этих  $n$  испытаниях. Очевидно, что эта случайная величина может принимать значения от 0 до  $n$ . Вероятность того, что она примет значение  $k$  ( $0 \leq k \leq n$ ) рассчитывается по формуле Бернулли:

$$P_n(k) = C_n^k p^k q^{n-k}, \quad (1.6)$$

где  $q = 1 - p$  – вероятность противоположного события, т. е. вероятность того, что  $A$  не наступит в результате испытания;

$$C_n^k = \frac{n!}{k!(n-k)!} \text{ – число различных сочетаний из } n \text{ по } k \text{ элементов.}$$

Математическое ожидание и дисперсия биномиального распределения рассчитываются по формулам:

$$M(X) = np; \quad D(X) = npq.$$

В MS Excel для расчета вероятности в задачах с фиксированным числом тестов или испытаний, когда результатом любого испытания может быть только успех или неудача, используется функция БИНОМРАСП(*число\_успехов*; *число\_испытаний*; *вероятность\_успеха*; *интегральная*). Здесь *число\_успехов* =  $k$ , *число\_испытаний* =  $n$ , *вероятность\_успеха* =  $p$ , а *интегральная* – это логическая величина, которая должна быть установлена равной 1, если необходимо получить функцию распределения, и равной 0, если нужно получить вероятность того, что случайная величина примет значение  $k$ .

**Пример 1.12.** Монета брошена 2 раза. Записать закон распределения случайной величины  $X$  – числа выпадений герба.

*Решение*

Вероятность появления герба в каждом бросании монеты равна:  $p = 1/2$ . Следовательно, вероятность выпадения надписи будет следующей:  $q = 1 - 1/2 = 1/2$ .

При двух бросаниях монеты герб может выпасть либо 2 раза, либо 1 раз, либо ни разу. Таким образом, значения  $X$  таковы:  $x_1 = 0$ ,  $x_2 = 1$ ,  $x_3 = 2$ . Каждое бросание монеты можно рассматривать как испытание, а искомую случайную величину – число появлений события «выпадение герба» в этих испытаниях. Следовательно, имеем биномиальный закон распределения.

Вероятности значений случайной величины  $X$  рассчитываются по формуле Бернулли следующим образом:

$$P_2(0) = C_2^0 p^0 q^{2-0} = \frac{2!}{0!2!} \cdot \left(\frac{1}{2}\right)^2 = 0,25;$$

$$P_2(1) = C_2^1 p^1 q^{2-1} = \frac{2!}{1!(2-1)!} \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = 0,5;$$

$$P_2(2) = C_2^2 p^2 q^{2-2} = \frac{2!}{2!(2-2)!} \cdot \left(\frac{1}{2}\right)^2 = 0,25.$$

Если использовать пакет MS Excel для расчетов, например  $P_2(1)$ , то с помощью *Мастера функций* нужно ввести функцию БИНОМРАСП(1;2;0,5;0).

### 1.3.2. Распределение Пуассона

Пусть производится  $n$  независимых испытаний, в каждом из которых вероятность появления события  $A$  равна  $p$ . При этом количество испытаний  $n$  очень велико, а вероятность  $p$  очень мала. В этом случае формула Бернулли не подходит, а используется асимптотическая формула Пуассона. Эта формула выведена при допущении, что произведение  $np$  является постоянной величиной, т. е.  $np = \lambda$ . Тогда вероятность того, что событие  $A$  наступит ровно  $k$  раз ( $k = 0, 1, 2 \dots$ ), определяется по формуле Пуассона:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \quad (1.7)$$

При этом параметр  $\lambda > 0$  называется *интенсивностью*, а закон распределения называется законом Пуассона вероятностей массовых ( $n$  – велико) и редких ( $p$  – мало) событий.

Дискретное распределение Пуассона проявляется в ситуациях, когда в течение определенного отрезка времени или на определенном пространстве происходит случайное число каких-либо событий (число радиоактивных распадов, случаи заболевания, число покупателей, проходящих в магазин за час и т. п.). Особенностью этого распределения является то, что дисперсия и математическое ожидание равны между собой:

$$M(X) = \lambda; \quad D(X) = \lambda.$$

В MS Excel используется функция ПУАССОН(*x*; *среднее*; *интегральный*). Она позволяет вычислить вероятность числа  $x$  появлений определенного события при заданном значении среднего (параметр  $\lambda$ ). При этом параметр *интегральный* должен быть равен 0. Если же он равен 1, то результатом является значение интегральной функции распределения.

**Пример 1.13.** На базу отправлено 10 000 изделий. Вероятность того, что изделие получит повреждение, равна 0,0003. Найти вероятность того, что на базу прибудет 4 поврежденных изделия.

*Решение*

По условию задачи  $n = 10\,000$ ,  $p = 0,0003$ ,  $k = 4$ . Рассчитываем параметр распределения:  $\lambda = n \cdot p = 10\,000 \cdot 0,0003 = 3$ . Используя формулу закона Пуассона, получаем результат:  $P_{10000}(X = 4) = \frac{3^4}{4!} \cdot e^{-3} = 0,168$ .

В приложении MS Excel эти расчеты можно было выполнить с помощью функции ПУАССОН(4;3;0). Параметр функции *интегральный* должен быть установлен в нуле, так как нам нужна вероятность, а не функция распределения.

### 1.3.3. Равномерное распределение

Равномерный закон распределения может быть как дискретным, так и непрерывным.

В случае дискретного равномерного распределения вероятность каждого из  $N$  возможных значений случайной величины одинакова:

$$P(x) = \frac{1}{N}. \quad (1.8)$$

Непрерывное распределение называется *равномерным*, если на интервале  $[a, b]$ , которому принадлежат все возможные значения случайной величины, плотность распределения имеет постоянное значение:

$$f(x) = \begin{cases} 0 & \text{при } x < a, \\ C & \text{при } a \leq x \leq b, \\ 0 & \text{при } x > b. \end{cases}$$

Поскольку площадь под функцией  $f(x)$  должна быть равна 1, можно определить значение константы ( $C$ ):

$$\int_a^b f(x) dx = \int_a^b C dx = Cx \Big|_a^b = C \cdot (b - a) = 1 \rightarrow C = \frac{1}{b - a}.$$

График функции плотности равномерного распределения имеет вид, показанный на рисунке 1.5.

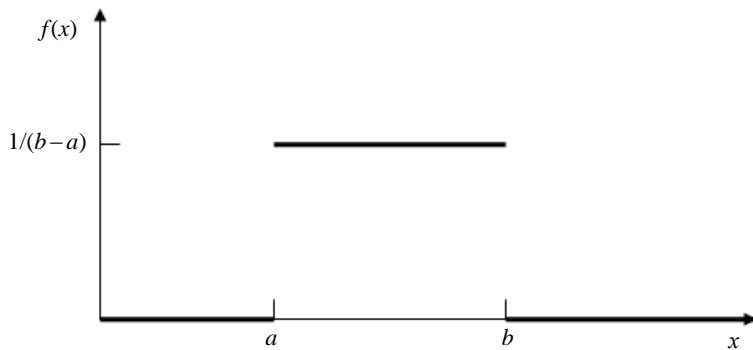


Рисунок 1.5 – График плотности равномерного распределения

Математическое ожидание и дисперсия равномерного распределения рассчитываются по формулам:

$$M(X) = \frac{a+b}{2}; \quad D(X) = \frac{(b-a)^2}{12}.$$

**Пример 1.14.** Непрерывная случайная величина имеет равномерный закон распределения на отрезке  $[2, 10]$ . Найти статистические характеристики этой случайной величины.

*Решение*

Найдем значение константы:  $C = \frac{1}{10-2} = \frac{1}{8}$ . Функция плотности распределения задается формулой

$$f(x) = \begin{cases} 0 & \text{при } x \leq 2, \\ \frac{1}{8} & \text{при } 2 < x \leq 10, \\ 0 & \text{при } x > 10. \end{cases}$$

Найдем значение математического ожидания, дисперсии и среднего квадратического отклонения:

$$M(X) = \frac{2+10}{2} = 6;$$

$$D(X) = \frac{(10-2)^2}{12} = \frac{64}{12} \approx 5,33;$$

$$\sigma(X) = \sqrt{D(X)} = \sqrt{5,33} = 2,3.$$

### 1.3.4. Нормальное распределение

Нормальное распределение (распределение Гаусса) широко распространено при описании случайных явлений, в которых на результат воздействует большое количество независимых случайных факторов, среди которых нет сильно выделяющихся. Например, рассеяние снарядов при стрельбе, распределение студентов по весу, ошибки измерения производственных параметров и т. п. Кроме того, при изучении явлений с помощью случайной выборки во многих случаях при большом объеме выборки закон распределения можно считать нормальным.

Плотность распределения вероятностей нормально распределенной случайной величины задается формулой

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-a)^2 / 2\sigma^2}, \quad (1.9)$$

где  $a$  и  $\sigma$  – параметры распределения (иногда используется краткое обозначение  $N(a, \sigma)$ ).

Можно показать, что параметр  $a$  есть математическое ожидание, а параметр  $\sigma$  – среднее квадратическое отклонение нормального распределения:

$$M(X) = a; \quad D(X) = \sigma^2; \quad \sigma(X) = \sigma.$$

График плотности распределения нормальной случайной величины показан на рисунке 1.6.

Таким образом, для нормального распределения крайние значения случайной величины (наименьшие и наибольшие) встречаются очень редко. Чем ближе значение к математическому ожиданию, тем чаще оно встречается. Функция  $f(x)$  симметрична относительно математического ожидания, т. е. отклонения в обе стороны равновероятны. Медиана и мода нормального распределения совпадают с математическим ожиданием.

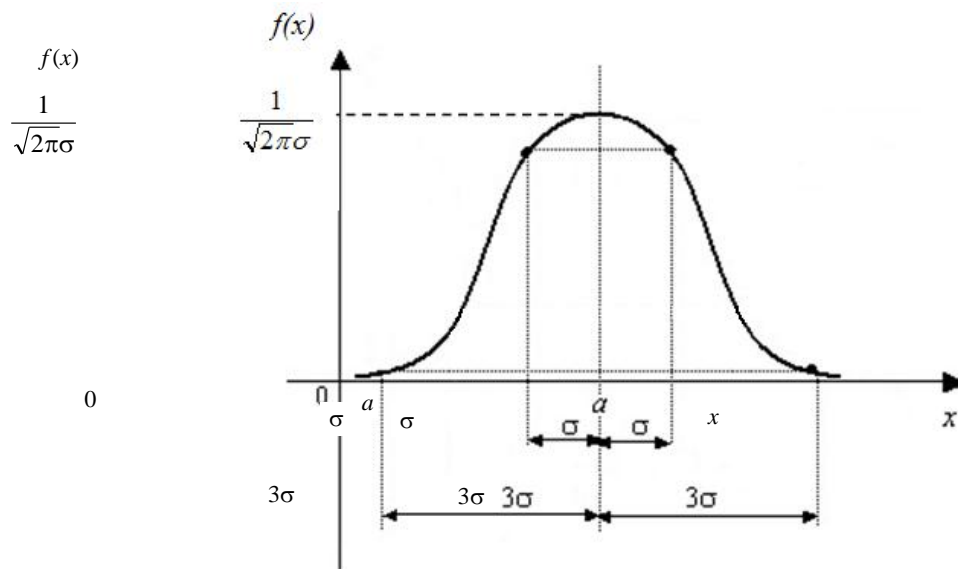


Рисунок 1.6 – Плотность распределения нормальной случайной величины

Параметр  $\sigma$  характеризует степень сжатия или растяжения диаграммы. Чем больше  $\sigma$ , тем шире и ниже кривая (рисунок 1.7).

В область от  $a - \sigma$  до  $a + \sigma$  нормально распределенная случайная величина попадает с вероятностью 0,683. В пределы от  $a - 2\sigma$  до  $a + 2\sigma$  случайная величина попадает с вероятностью 0,955, а в пределы от  $a - 3\sigma$  до  $a + 3\sigma$  – с вероятностью 0,997. Последняя закономерность называется правилом «трех сигм».

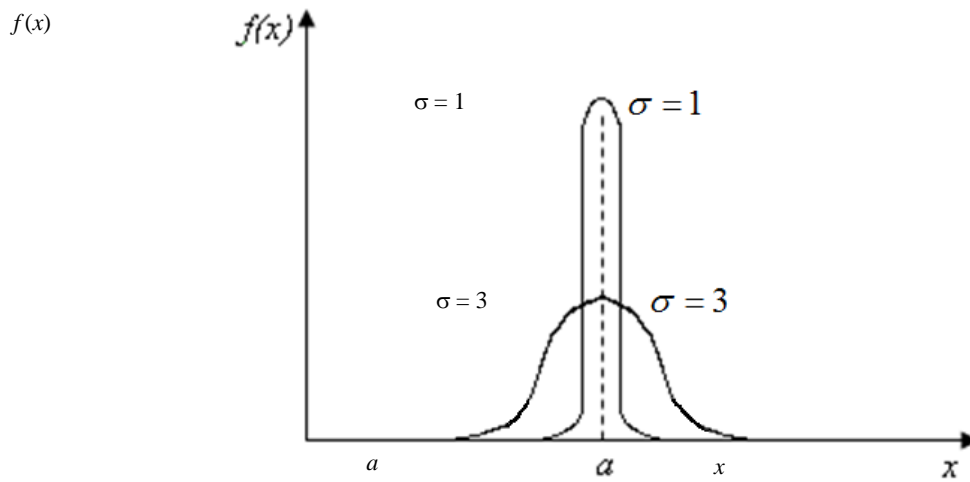


Рисунок 1.7 – Влияние значения среднего квадратического отклонения на форму нормальной кривой

Особую роль играет нормальное распределение с параметрами  $a = 0$  и  $\sigma = 1$ , т. е. распределение  $N(0,1)$ , которое называют *стандартным* или *нормированным нормальным распределением*. Соответствующая функция плотности распределения имеет следующий вид:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad (1.10)$$

В Excel для вычисления значений нормального распределения используются следующие функции:

1. НОРМРАСП( $x$ ; *среднее*; *стандартное откл*; *интегральная*) – вычисление значения функции плотности распределения или функции распределения нормальной случайной величины, при этом:

- $x$  – значение случайной величины, для которого вычисляется функция;
- *среднее* – математическое ожидание (параметр  $a$ );
- *стандартное откл* – стандартное отклонение (параметр  $\sigma$ );
- *интегральная* – логическое значение, определяющее форму функции (если ИСТИНА, то строится интегральная функция распределения, а если ЛОЖЬ – функция плотности распределения).

2. НОРМСТРАСП( $x$ ) – вычисление значения функции распределения стандартной нормальной случайной величины для аргумента  $x$ .

**Пример 1.15.** Построить график функции плотности нормального распределения  $f(x)$  при  $a = 24,3$  и  $\sigma = 1,5$ . Рассчитать вероятность попадания этой случайной величины в интервал  $[25,27)$ .

*Решение*

В ячейку A1 введем символ случайной величины ( $X$ ), а в ячейку B1 – обозначение плотности распределения ( $f(x)$ ). В ячейки D2 и E2 внесем исходные данные распределения – значения  $a$  и  $\sigma$  (рисунок 1.8).

Чтобы рассчитать диапазон изменения значений  $x$ , воспользуемся правилом «трех сигм», согласно которому 99,7% всех значений величин попадает в интервал  $a \pm 3\sigma$ . В ячейке F2 рассчитаем нижнюю границу этого диапазона, введя формулу  $=D2-3*E2$ . В ячейке G2 аналогично рассчитаем верхнюю границу диапазона. В диапазон A2:A20 введем значения  $x$  от 19,8 до 28,8 с шагом 0,5.

Установим табличный курсор в ячейку B2, введем с помощью *Мастера функций* стандартную функцию из категории *Статистические* НОРМРАСП() (рисунок 1.9).

Абсолютные ссылки на такие аргументы этой функции, как среднее и стандартное отклонение, поставлены для того, чтобы правильно скопировать эту формулу для всех значений  $x$ . Выполним это копирование методом автозаполнения. В результате получим значения функции плотности распределения, которые показаны на рисунке 1.8. По этим данным построим диаграмму типа «точечная, соединенная отрезками».

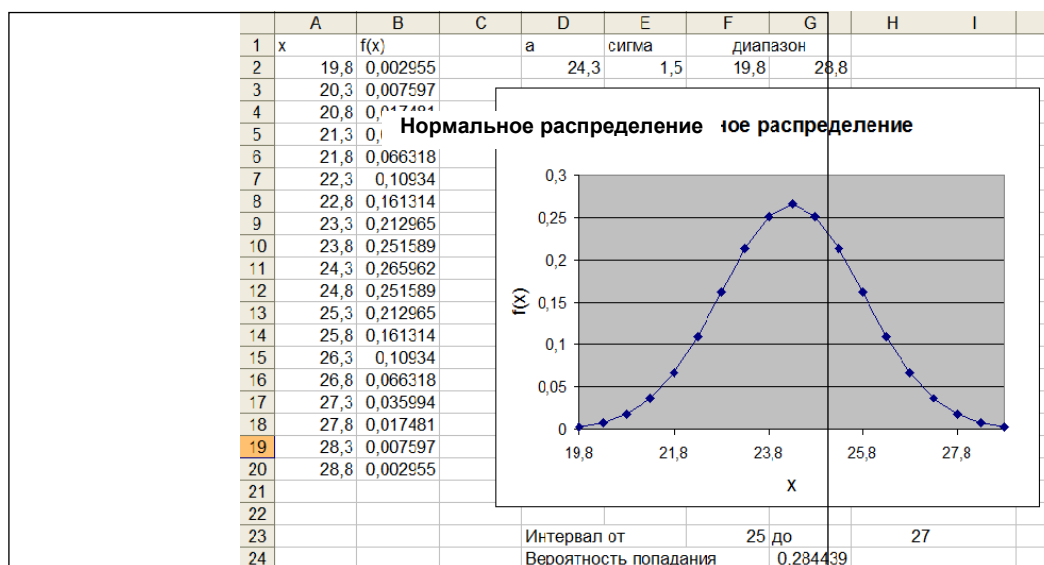


Рисунок 1.8 – График функции плотности нормального распределения в Excel

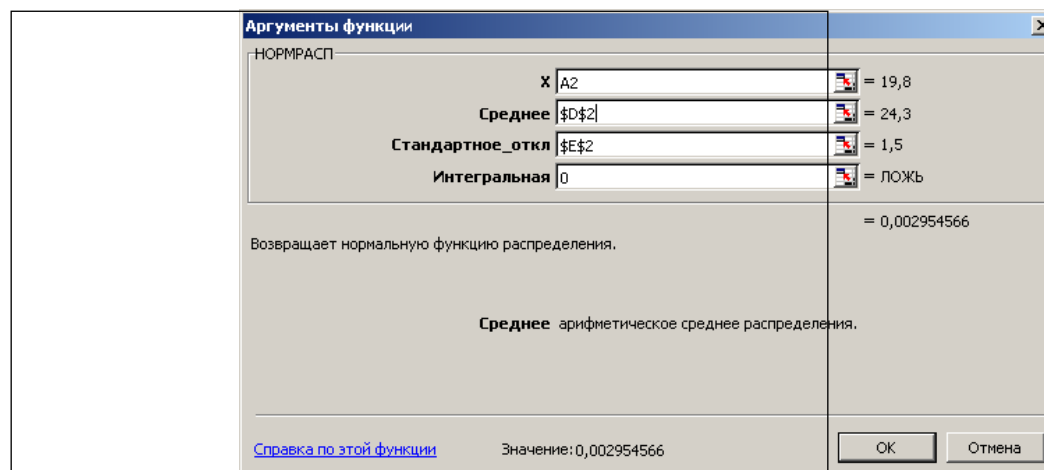


Рисунок 1.9 – Окно Мастера функций для ввода функции НОРМРАСП()

Для расчета вероятности попадания в заданный интервал воспользуемся третьим свойством функции распределения:  $P(25 \leq x < 27) = F(27) - F(25)$ . Значения функции распределения легко получить с помощью той же функции Excel НОРМРАСП(), если аргумент *интегральная* установить равным 1. В ячейки F23 и H23 занесем границы искомого интервала (см. рисунок 1.8), а в ячейку G24 – формулу для расчета вероятности попадания в интервал:

$$=НОРМРАСП(H23;D2;E2;1)-НОРМРАСП(F23;D2;E2;1).$$

Как видно из рисунка 1.8, результат расчета по этой формуле составляет приблизительно 0,28, т. е. 28% всех значений данной случайной величины приходятся на интервал от 25 до 27.

### 1.3.5. Экспоненциальное (показательное) распределение

Экспоненциальное распределение – это непрерывное распределение, которое обычно используется в задачах массового обслуживания. Например, распределение интервалов между обращением граждан во многие организации, время обслуживания одного запроса и т. д. Плотность экспоненциального распределения задана функцией

$$f(x) = \begin{cases} 0 & \text{при } x < 0, \\ \lambda \cdot e^{-\lambda x} & \text{при } x \geq 0, \end{cases}$$

где  $\lambda > 0$  – параметр распределения.

График плотности экспоненциального распределения показан на рисунке 1.10.



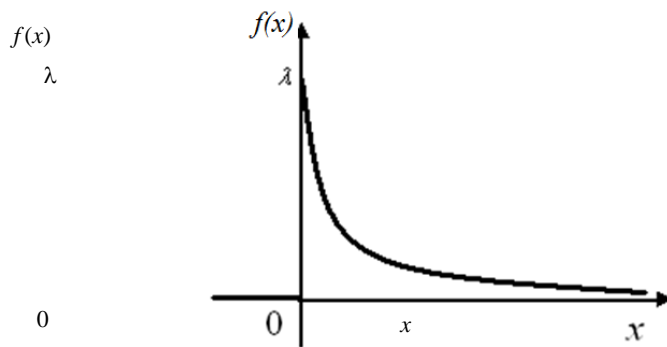


Рисунок 1.10 – Функция плотности экспоненциального (показательного) распределения

Математическое ожидание и дисперсия рассчитываются по следующим формулам:

$$M(X) = \frac{1}{\lambda}; \quad D(X) = \frac{1}{\lambda^2}.$$

В Excel для вычисления значений экспоненциального распределения используется функция ЭКСПРАСП( $x$ ;  $\lambda$ ;  $\text{интегральная}$ ). Она позволяет вычислить вероятность (плотность распределения) или значение функции распределения (зависит от аргумента  $\text{интегральная}$ ) для заданного значения  $x$ .

**Пример 1.16.** Построить график функции плотности распределения экспоненциального закона, если известно, что математическое ожидание равно 5.

*Решение*

Определим сначала параметр распределения:  $\lambda = \frac{1}{M(X)} = \frac{1}{5} = 0,2$ . Занесем это значение в ячейку E2.

В столбец A введем значения аргумента  $x \geq 0$  с шагом 0,5 (рисунок 1.11). Поскольку с увеличением  $x$  график приближается к нулю, количество элементов этого столбца значения не имеет. В ячейку B2 введем формулу ЭКСПРАСП(A2;E\$2;0), которую скопируем вниз по столбцу методом автозаполнения. Построим по этим данным диаграмму типа «точечная, соединенная отрезками» (рисунок 1.11).

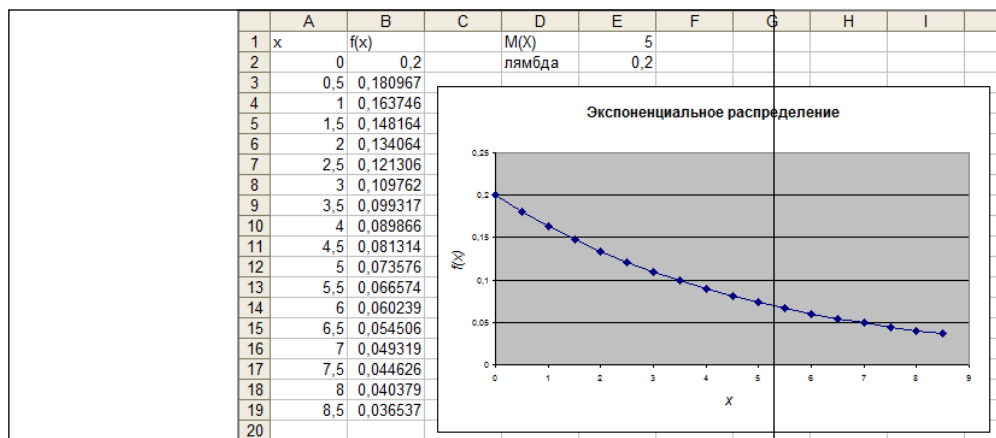


Рисунок 1.11 – График функции плотности экспоненциального распределения в Excel

### 1.3.6. Распределение $\chi^2$ (хи-квадрат)

Если случайные величины  $Y_1, Y_2, \dots, Y_n$  независимы, и каждая из них имеет стандартное нормальное распределение  $N(0,1)$ , то случайная величина  $\chi_n^2 = Y_1^2 + Y_2^2 + \dots + Y_n^2$  имеет распределение хи-квадрат с  $n$  степенями свободы. Другое название этого распределения – распределение Пирсона.

График плотности вероятности случайной величины, имеющей  $\chi^2$ -распределение, лежит только в первой четверти декартовой системы координат и имеет асимметричный вид с вытянутым правым «хвостом». Однако с увеличением числа степеней свободы распределение  $\chi^2$  постепенно приближается к нормальному (рисунок 1.12).

Математическое ожидание и дисперсия этой случайной величины рассчитываются по формулам:

$$M(\chi_n^2) = n; \quad D(\chi_n^2) = 2n.$$

Распределение  $\chi^2$  применяется для нахождения интервальных оценок и проверки статистических гипотез.

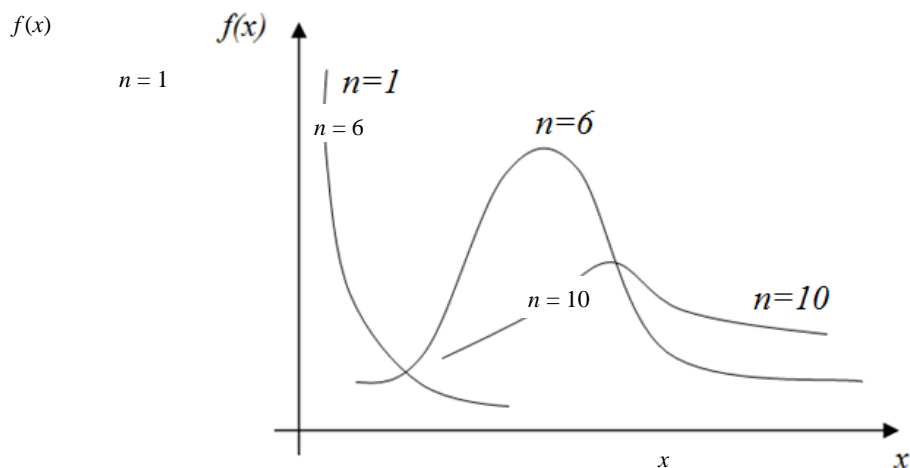


Рисунок 1.12 – Плотность распределения  $\chi^2$

В Excel для вычисления значений распределения хи-квадрат используются следующие функции:

1. ХИ2РАСП( $x$ ; *степени\_свободы*) – расчет вероятности того, что значение случайной величины будет больше заданного значения  $x$  (ХИ2РАСП= $P(X > x)$ ).
2. ХИ2ОБР(*вероятность*; *степени\_свободы*) – расчет значения  $x$ , для которого  $P(X > x) = \text{вероятность}$ . Эта функция – обратная функции ХИ2РАСП().

### 1.3.7. Распределение Стьюдента

Если случайные величины  $Y_0, Y_1, Y_2, \dots, Y_n$  независимы, и каждая из них имеет стандартное нормальное распределение  $N(0,1)$ , то случайная величина

$$t_n = \frac{Y_0}{\sqrt{\frac{1}{n} \sum_{i=1}^n Y_i^2}} \quad (1.11)$$

имеет распределение, называемое распределением Стьюдента с  $n$  степенями свободы.

Математическое ожидание и дисперсия этой случайной величины равны:

$$M(t_n) = 0; \quad D(t_n) = \frac{n}{n-2} \quad (\text{существует только при } n > 2).$$

График функции плотности вероятности случайной величины, имеющей распределение Стьюдента, является симметричной кривой (линия симметрии – ось ординат) (рисунок 1.13). При увеличении числа степеней свободы распределение Стьюдента приближается к стандартному нормальному, причем при  $n > 30$  распределение Стьюдента практически можно заменить нормальным распределением.

Распределение Стьюдента применяется для нахождения интервальных оценок и проверки статистических гипотез.

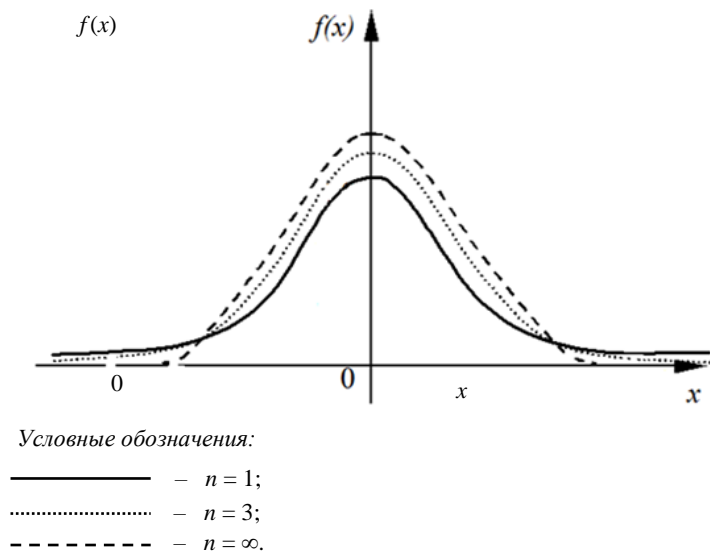


Рисунок 1.13 – Графики плотности распределения Стюдента при различных степенях свободы

В Excel для вычисления значений распределения Стюдента используются следующие функции:

1. СТЬЮДРАСП( $x$ ;  $\text{степени\_свободы}$ ;  $\text{хвосты}$ ) – расчет вероятности по заданному значению  $x$ . Аргумент  $\text{хвосты}$  может принимать значения 1 (одностороннее распределение) или 2 (двустороннее распределение). Если  $\text{хвосты} = 1$ , то функция СТЬЮДРАСП вычисляется как  $\text{СТЮДРАСП} = P(X > x)$ , где  $x$  – это случайная величина, соответствующая  $t$ -распределению. Если  $\text{хвосты} = 2$ , то функция СТЬЮДРАСП вычисляется как  $\text{СТЮДРАСП} = P(|X| > x) = P(X > x \text{ или } X < -x)$ . Данную функцию можно использовать вместо таблицы критических значений распределения Стюдента.

2. СТЬЮДРАСПОБР( $\text{вероятность}$ ;  $\text{степени\_свободы}$ ) – расчет значения  $x$  по заданной вероятности для двустороннего распределения. СТЬЮДРАСПОБР возвращает значение  $x$ , для которого  $P(|X| > x) = \text{вероятность}$ . Для одностороннего распределения можно заменить аргумент  $\text{вероятность}$  на  $2 * \text{вероятность}$ .

### 1.3.8. Распределение Фишера

Если случайные величины  $Y_1, Y_2, \dots, Y_n$  и  $Z_1, Z_2, \dots, Z_m$  независимы, и каждая из них имеет стандартное нормальное распределение  $N(0,1)$ , где  $n$  и  $m$  – натуральные числа, то случайная величина

$$F_{m,n} = \frac{\frac{1}{m}(Z_1^2 + Z_2^2 + \dots + Z_m^2)}{\frac{1}{n}(Y_1^2 + Y_2^2 + \dots + Y_n^2)} \quad (1.12)$$

имеет распределение Фишера (или  $F$ -распределение) с параметрами  $m$  и  $n$ , называемыми степенями свободы. Таким образом, распределение Фишера определяется двумя параметрами (числами степеней свободы).

При больших значениях  $m$  и  $n$  это распределение приближается к нормальному. Очевидно, что  $F_{1,n}$  – это не что иное, как распределение Стюдента с  $n$  степенями свободы. Графики функции плотности распределения Фишера для различных параметров  $m$  и  $n$  показаны на рисунке 1.14.

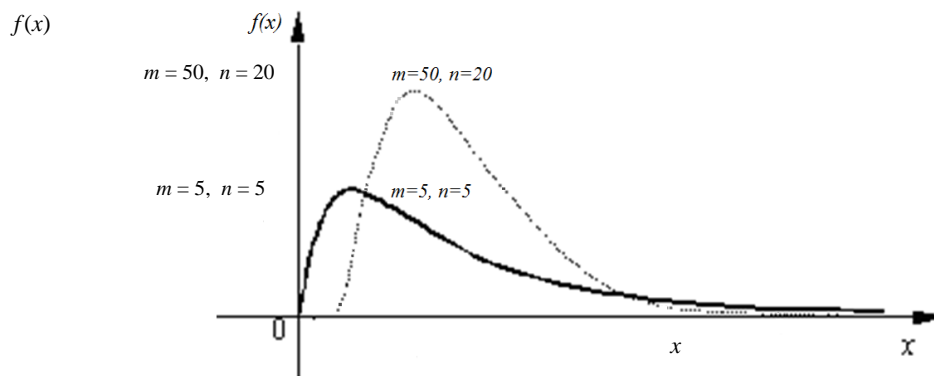


Рисунок 1.14 – Графики плотности распределения Фишера при различных степенях свободы

Математическое ожидание и дисперсия для  $F$ -распределения рассчитываются по формулам:

$$M(F_{m,n}) = \frac{n}{n-2} \quad (\text{существует при } n > 2);$$

$$D(F_{m,n}) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (\text{существует при } n > 4).$$

Распределение Фишера используется при проверке статистических гипотез, в дисперсионном и регрессионном анализе.

В Excel для вычисления значений распределения Фишера используются следующие функции:

1. **ФРАСП**( $x$ ; *степени\_свободы1*; *степени\_свободы2*) – расчет вероятности того, что значение случайной величины будет больше заданного значения  $x$  ( $\text{ФРАСП} = P(X > x)$ ).

2. **ФРАСПОБР**(*вероятность*; *степени\_свободы1*; *степени\_свободы2*) – расчет значения  $x$  по заданной вероятности (функция, обратная функции **ФРАСП**).

### Задания для самостоятельной работы\*

1. Банк выдает 5 кредитов. Вероятность невозвращения кредита равна 0,2 для каждого из заемщиков. Составьте таблицу закона распределения количества заемщиков, не вернувших кредит по окончании срока кредитования. Постройте многоугольник распределения и функцию распределения. Рассчитайте числовые характеристики.

2. Игральная кость брошена 3 раза. Напишите закон распределения числа появлений шестерки. Постройте многоугольник распределения и функцию распределения. Рассчитайте числовые характеристики.

3. Пряжильщица обслуживает 1000 веретен. Вероятность обрыва нити на одном веретене в течение 1 мин равна 0,004. Найдите вероятность того, что в течение 1 мин обрыв произойдет на пяти веретенах.

4. Коммутатор учреждения обслуживает 100 абонентов. Вероятность того, что в течение 1 мин абонент позвонит на коммутатор, равна 0,02. Определите, какое из двух событий вероятнее:

- в течение 1 мин позвонят 3 абонента;
- в течение 1 мин позвонят 4 абонента.

5. Рукопись объемом в 1000 страниц машинописного текста содержит ровно 1000 опечаток. Найдите вероятность того, что наудачу взятая страница содержит ровно 2 опечатки. Предполагается, что число опечаток распределено по закону Пуассона.

6. Магазин производит продажу мужских костюмов. По данным статистики распределение по размерам является нормальным с математическим ожиданием и средним квадратическим отклонением, соответственно равным 48 и 2. Определите процент спроса на 50-й размер при условии разброса значений этой величины в интервале (49,51).

7. Валики, изготавливаемые автоматом, считаются стандартными, если отклонение диаметра валика от проектного размера не превышает 2 мм. Случайные отклонения диаметра валиков подчиняются нормаль-

\* При решении можно использовать MS Excel.

ному закону распределения со средним квадратическим отклонением  $\sigma = 1,6$  мм и математическим ожиданием  $a = 0$ . Постройте график функции плотности распределения и определите, сколько процентов стандартных валиков изготавливает автомат.

8. Определите математическое ожидание и среднее квадратическое отклонение непрерывной случайной величины, которая имеет равномерный закон распределения на интервале  $[4, 10]$ . Постройте график функции плотности распределения.

9. Непрерывная случайная величина  $X$  имеет экспоненциальный закон распределения с параметром  $\lambda = 10$ . Найдите математическое ожидание и дисперсию величины  $X$ . Постройте график функции плотности распределения. Определите вероятность того, что в результате испытания значение случайной величины  $X$  попадет в интервал  $(0,2; 0,5)$ .

## 2. ВЫБОРКА И ЕЕ АНАЛИЗ

### 2.1. Построение и визуализация вариационного ряда

В разделе 1 настоящего пособия рассматривались различные законы распределения случайных величин. При этом предполагалось, что закон распределения задан тем или иным способом (таблицей, функцией распределения, функцией плотности распределения и др.). Однако для тех случайных величин, с которыми мы имеем дело в реальной жизни, закон распределения, как правило, неизвестен. Задача исследователя – на основе анализа данных наблюдений или экспериментов выяснить, каков этот закон, определить его параметры. Месячный объем продаж определенного товара в различных магазинах, среднедушевой доход семьи, геометрические размеры производимых деталей – все это примеры случайных величин. Судить об их законах распределения можно только на основании наблюдений, которые, как правило, охватывают только часть возможных значений этих величин.

Математическая статистика занимается разработкой методов сбора и обработки данных, получаемых в результате наблюдений массовых случайных явлений. Методы сбора статистических данных останутся за рамками данного пособия. С ними можно ознакомиться в литературе по теории вероятностей и математической статистике. Основное внимание будет уделено методам анализа данных наблюдений для получения научных и практических выводов.

Исследование случайной величины может быть сплошным и выборочным. При сплошном исследовании рассматриваются все возможные значения случайной величины. Такой подход применяют крайне редко, поскольку это связано с большим объемом работ, а часто просто нереально.

*Выборочный метод* исследования состоит в том, что отбираются некоторые значения случайной величины, на основе анализа которых делается вывод относительно всей совокупности возможных значений (так повар по одной ложке делает вывод о содержимом всей кастрюли). Вся совокупность значений случайной величины называется *генеральной совокупностью*. Группа случайно отобранных значений из генеральной совокупности называется *выборочной совокупностью (выборкой)*. Число элементов выборки называется ее *объемом*. Выборка должна наилучшим образом представлять всю генеральную совокупность, т. е. быть *репрезентативной (представительной)*. Для этого нужно четко определить, что понимается под генеральной совокупностью. Например, если нужно оценить количество поступающих из сельской местности во все вузы города, то абитуриенты одного университета – это нерепрезентативная выборка. Когда о генеральной совокупности недостаточно сведений, обычно выборка формируется случайным отбором.

Пусть из генеральной совокупности извлечена выборка объемом  $n$ , причем известно, что исследуемая случайная величина является дискретной. В этой выборке значение  $x_1$  наблюдалось  $n_1$  раз, значение  $x_2$  –  $n_2$  раз, значение  $x_k$  –  $n_k$  раз. Значения  $x_i$  называются *вариантами*, соответствующие им значения  $n_i$  ( $i = 1, 2, \dots, k$ ) называются *частотами*. Отношение частоты к объему выборки называется *относительной частотой*.

Относительная частота выражается формулой

$$W_i = n_i / n,$$

где  $\sum_{i=1}^k n_i = n$  и  $\sum_{i=1}^k W_i = 1$ .

Перечень вариантов в возрастающем порядке и соответствующих им частот или относительных частот называется *вариационным рядом* (таблица 2.1). Он является аналогом закона распределения дискретной случайной величины, с той разницей, что относительная частота – это лишь оценка вероятности, полученная по выборке.

Таблица 2.1 – Вариационный ряд

Варианты	$x_1$	$x_2$	...	$x_k$
Частота	$n_1$	$n_2$	...	$n_k$
Относительная частота	$W_1$	$W_2$	...	$W_k$

Если каждую пару  $(x_i, n_i)$  изобразить точкой на координатной плоскости и соединить эти точки ломаной линией, то будет получен *полигон частот*. Ломаная, соединяющая на координатной плоскости точки  $(x_i, W_i)$ , называется *полигоном относительных частот*. Это аналог многоугольника распределения.

Пусть теперь извлечена выборка объема  $n$  из генеральной совокупности, для которой известно, что исследуемая случайная величина является непрерывной. Весь диапазон изменения наблюдаемых значений этой случайной величины  $[x_{\min}, x_{\max}]$  разбивается на ряд интервалов одинаковой ширины  $h$ . Число интервалов обычно выбирают не менее 5 и не более 15 (число интервалов – целая часть числа, определяемого по одной из формул:  $m \approx \sqrt{n}$ ,  $m \approx 5 \lg n$ ,  $m \approx 1 + 3,322 \lg n$ ). Затем определяется число значений случайной величины, попавших в каждый интервал ( $n_i$ ). Если значение случайной величины попадает на границу интервалов, то она учитывается в левом интервале (кроме минимального значения, которое учитывается в первом интервале). Относительная частота попадания в заданный интервал рассчитывается по формуле  $W_i = n_i / n$ . Полученный вариационный ряд называется *интервальным* (таблица 2.2). Аналогично поступают и с дискретной случайной величиной, если число ее вариантов велико (больше 15).

Таблица 2.2 – Интервальный вариационный ряд

Варианты	$[x_0, x_1]$	$(x_1, x_2]$	...	$(x_{k-1}, x_k]$
Частота	$n_1$	$n_2$	...	$n_k$
Относительная частота	$W_1$	$W_2$	...	$W_k$

По этим данным можно построить *гистограммы частот* и *относительных частот*. В математической статистике гистограммой называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы шириной  $h$  и высотой, равной  $n_i/h$  (для частот) или  $W_i/h$  (для относительных частот). Таким образом, площадь каждого прямоугольника равна частоте (относительной частоте). Однако в Excel вид гистограммы несколько упрощен: ширина каждого прямоугольника не зависит от размера частичного интервала  $h$ , а высота равна частоте или относительной частоте.

Гистограмма относительных частот является аналогом функции плотности распределения случайной величины.

Иногда строится также *отсортированная гистограмма* (по Парето). В ней прямоугольники располагаются в порядке убывания частот (или относительных частот).

*Выборочной (эмпирической) функцией распределения* называется функция  $F_n(x) = n_x / n$ , где  $n_x$  – число значений случайной величины, меньших  $x$ ;  $n$  – объем выборки.

При большом числе наблюдений эмпирическая функция распределения  $F_n(x)$  приближается к теоретической интегральной функции распределения генеральной совокупности  $F(x)$ .

Эмпирическая функция распределения обладает всеми свойствами теоретической функции:

- ее значения принадлежат отрезку  $[0, 1]$ ;
- она является неубывающей функцией;
- $F_n(x) = 0$  при  $x \leq x_{\min}$  и  $F_n(x) = 1$  при  $x > x_{\max}$ .

Очевидно, что эмпирическая функция распределения имеет ступенчатый вид. Однако в Excel построить такую функцию сложно, поэтому ее заменяют ломаной линией, проходящей через левые концы «ступенек». Такая ломаная линия называется *кумулятивной кривой*.

**Пример 2.1.** Медпункт некоторого учебного заведения проводит измерения веса учащихся (в килограммах). Произвольным образом отобрали результаты 20 человек: 64, 57, 62, 58, 61, 63, 60, 61, 65, 62, 62, 60, 64, 61, 59, 59, 63, 61, 62, 58. Построить вариационный ряд по этой выборке, а также полигон относительных частот, эмпирическую функцию распределения и ее график.

#### Решение

По смыслу вес – величина непрерывная. Однако в данном случае вес всех учащихся измерен с точностью до килограмма. Поэтому можно рассматривать ее как дискретную. Минимальное наблюдавшееся значение веса ( $x_{\min}$ ) равно 57, максимальное значение ( $x_{\max}$ ) – 65. Всего наблюдений ( $n$ ) равно 20. Построим вариационный ряд по данной выборке (таблица 2.3). Чтобы определить частоты  $n_i$  ( $i = 1, 2 \dots 9$ ), подсчитаем, сколько раз каждая из девяти вариантов встречается в выборке (например, число 61 встречается в выборке четыре раза). Для определения относительных частот  $W_i$  ( $i = 1, 2 \dots 9$ ) разделим соответствующую частоту на объем выборки ( $n = 20$ ).

Полигон относительных частот показан на рисунке 2.1.

Таблица 2.3 – Вариационный ряд и накопленные частоты для примера 2.1

Варианты	57	58	59	60	61	62	63	64	65
Частота	1	2	2	2	4	4	2	2	1
Относительная частота	0,05	0,1	0,1	0,1	0,2	0,2	0,1	0,1	0,05
Накопленная частота	0,05	0,15 = = 0,05 + + 0,1	0,25 = = 0,15 + + 0,1	0,35 = = 0,25 + + 0,1	0,55 = = 0,35 + + 0,2	0,75 = = 0,55 + + 0,2	0,85 = = 0,75 + + 0,1	0,95 = = 0,85 + + 1	1 = = 0,95 + + 0,05

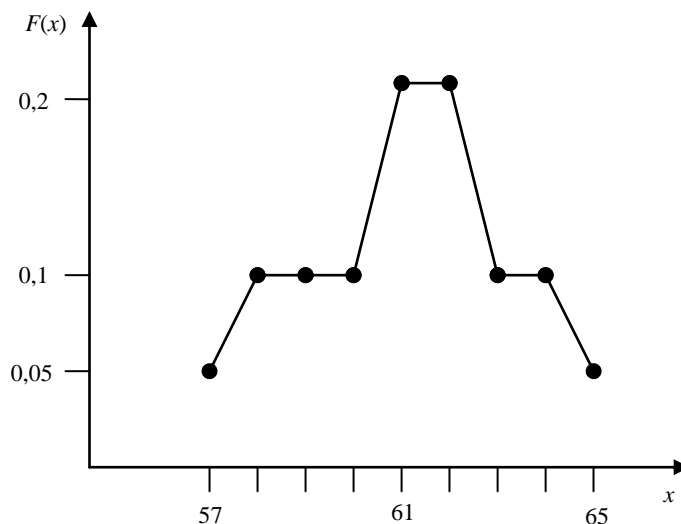


Рисунок 2.1 – Полигон относительных частот для примера 2.1

Для построения эмпирической функции распределения рассчитаем в таблице 2.3 накопленные частоты (суммы частот всех вариантов, меньших либо равных данной). Накопленная частота является высотой «ступеньки», а соответствующее значение варианты – ее левым концом (который не включается при записи эмпирической функции распределения). Таким образом, эмпирическая функция распределения имеет вид

$$F_{20}(x) = \begin{cases} 0 & \text{при } x \leq 57, \\ 0,05 & \text{при } 57 < x \leq 58, \\ 0,15 & \text{при } 58 < x \leq 59, \\ 0,25 & \text{при } 59 < x \leq 60, \\ 0,35 & \text{при } 60 < x \leq 61, \\ 0,55 & \text{при } 61 < x \leq 62, \\ 0,75 & \text{при } 62 < x \leq 63, \\ 0,85 & \text{при } 63 < x \leq 64, \\ 0,95 & \text{при } 64 < x \leq 65, \\ 1 & \text{при } x > 65. \end{cases}$$

Кумулятивная кривая, которая строится в MS Excel вместо ступенчатой функции распределения, проходит как раз по точкам с координатами  $(x_i, \text{накопленная частота для } x_i)$ . График эмпирической функции распределения и соответствующая кумулятивная кривая показаны на рисунке 2.2.

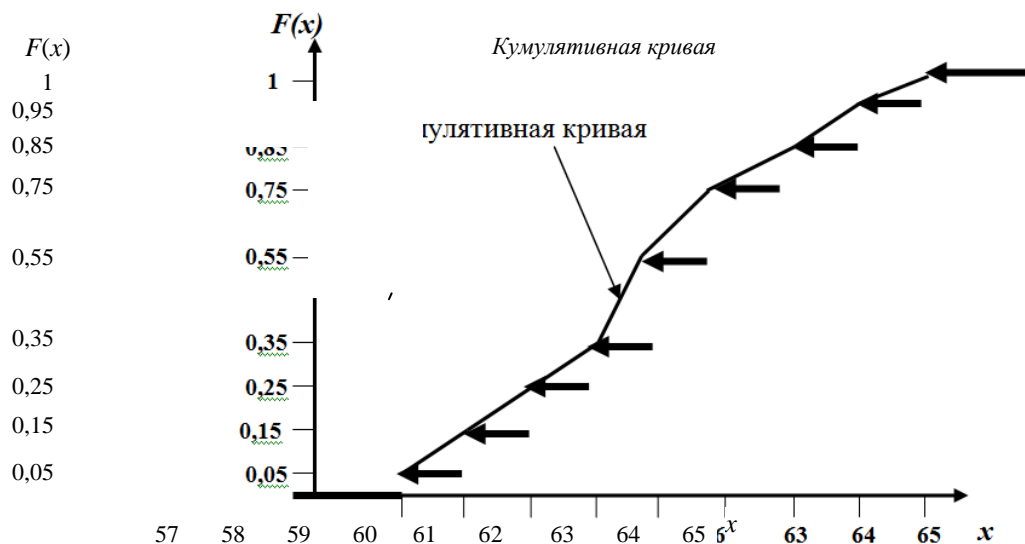


Рисунок 2.2 – Эмпирическая функция распределения и кумулятивная кривая

**Пример 2.2.** Построить интервальный вариационный ряд по выборке веса учащихся: 64, 57, 62, 58, 61, 63, 60, 61, 65, 62, 62, 60, 64, 61, 59, 59, 63, 61, 62, 58.

*Решение*

Определим количество интервалов, на которое следует разбить диапазон значений наблюдаемой случайной величины:  $\sqrt{20} \approx 4,47$ ;  $5 \lg 20 \approx 6,5$ ;  $1 + 3,322 \lg 20 \approx 5,62$ . Таким образом, разбивать диапазон значений следует на 4–6 интервалов. Если взять шаг 2, то получается 5 интервалов (таблица 2.4). Специфика подсчета частоты в программе MS Excel такова, что для каждого интервала подсчитывается число значений в выборке, *меньших* либо *равных* правой границы и *строго больших* левой. Чтобы минимальное значение по выборке ( $x_{\min}$ ) не оказалось единственным в первом интервале, желательно выбирать его левую границу меньше, чем  $x_{\min}$ . Тогда минимальное значение попадет внутрь первого интервала. Поскольку в данной выборке  $x_{\min} = 57$ , возьмем левую границу первого интервала, равную 56. Подсчитаем частоты – количество наблюдений, которые попадают в каждый интервал. Относительные частоты рассчитаем как отношение частот к объему выборки.

Таблица 2.4 – Интервальный вариационный ряд для примера 2.2

Варианты	[56,58]	(58,60]	(60,62]	(62,64]	(64,66]
Частота	3	4	8	4	1
Относительная частота	0,15	0,2	0,4	0,2	0,05
Накопленная частота	0,15	0,35	0,75	0,95	1

Для полученного интервального ряда построим классическую гистограмму (площадь каждого столбика равна относительной частоте) (рисунок 2.3).



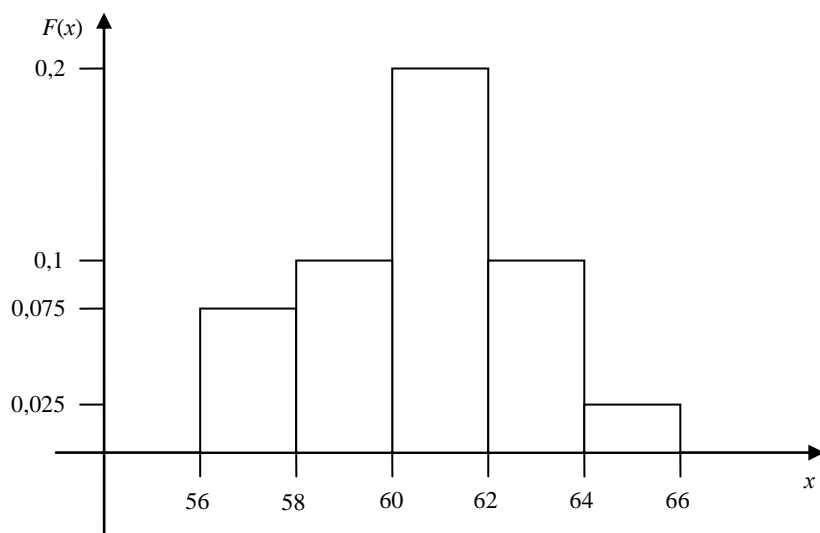


Рисунок 2.3 – Гистограмма для интервального вариационного ряда

В четвертой строке таблицы 2.4 рассчитаем накопленные частоты и построим эмпирическую функцию распределения:

$$F_{20}(x) = \begin{cases} 0 & \text{їдї} x \leq 58, \\ 0,15 & \text{їдї} 58 < x \leq 60, \\ 0,35 & \text{їдї} 60 < x \leq 62, \\ 0,75 & \text{їдї} 62 < x \leq 64, \\ 0,95 & \text{їдї} 64 < x \leq 66, \\ 1 & \text{їдї} x > 66. \end{cases}$$

График этой функции и соответствующей кумулятивной кривой показан на рисунке 2.4.

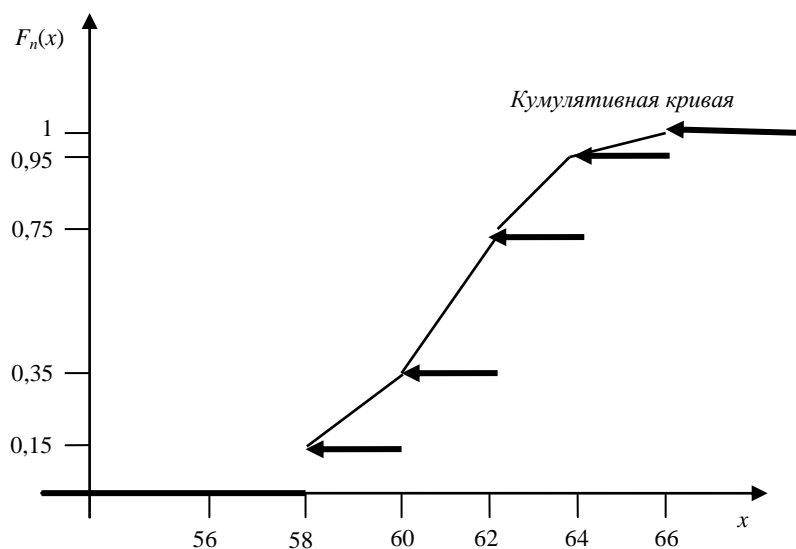


Рисунок 2.4 – Эмпирическая функция распределения и кумулятивная кривая

**Пример 2.3.** Дана выборка значений веса учащихся в килограммах. Объем этой выборки ( $n$ ) равен 55 (таблица 2.5). Построить вариационный ряд, используя MS Excel.

Таблица 2.5 – Эмпирические данные о весе учащихся

Наблюдения				
64	62	58	63	61
57	62	63	58	58
63	60	61	60	60
62	64	59	59	64
58	61	62	60	60
61	59	60	59	59
63	59	60	61	61
60	63	58	62	64
60	61	61	62	62
61	62	60	63	59
65	58	63	57	65

**Решение**

1. В ячейку A1 введем заголовок «Наблюдения», а в диапазон A2:A56 – эмпирические данные, приведенные в таблице 2.5.

2. Рассчитаем максимальное и минимальное значения выборочных данных в ячейках D1 и D2, введя соответственно функции МАКС(A2:A56) и МИН(A2:A56) (рисунок 2.5).

В этом примере построим вариационный ряд, считая вес дискретной случайной величиной. В ячейку E1 введем заголовок «Варианты», а ниже в столбце – все возможные неповторяющиеся значения веса учащихся ( $x_i$ ), которые встречались в выборке (от минимального до максимального).

3. В ячейке F1 запишем заголовок «Абсолютные частоты». В этом столбце будут рассчитаны значения частот  $n_i$ , т. е. то количество раз, которое соответствующее значение  $x_i$  случайной величины встречалось в выборке. Для заполнения столбца абсолютных частот можно использовать стандартную функцию ЧАСТОТА(). Выделим мышью диапазон F2:F10, в котором разместятся найденные частоты, вызовем *Мастер функций* и в категории *Статистические* выберем функцию ЧАСТОТА. После этого заполним ее аргументы:

- *Массив данных* – это диапазон эмпирических данных A2:A56.
- *Массив интервалов* – это диапазон значений вариант E2:E10.

Закончить ввод функции нужно одновременным нажатием клавиш <Ctrl> + <Shift> + <Enter>, поскольку ее результатом является диапазон значений. В строке формул эта функция будет показана в фигурных скобках.

В ячейке F11 найдем общее число наблюдений, просуммировав значения в столбце абсолютных частот (рисунок 2.5).

4. В ячейке G1 запишем заголовок «Относительные частоты». Для расчета относительных частот  $W_i$  внесем в ячейку G2 формулу =F2/\$F\$11 и скопируем ее методом автозаполнения вниз по столбцу. Сумма относительных частот в этом столбце должна быть равна единице.

5. Последний столбец таблицы озаглавим «Накопленные частоты». В ячейку H2 скопируем значение относительной частоты из ячейки G2, а в ячейку H3 введем формулу =H2+G3. Методом автозаполнения скопируем введенную формулу вниз по столбцу в диапазон H4:H10.

Итоговый вид таблицы после форматирования показан на рисунке 2.5.

	A	B	C	D	E	F	G	H
1	Наблюдения	Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты	
2	64	Минимум	57	57	2	0,036	0,036	
3	57			58	6	0,109	0,145	
4	63			59	7	0,127	0,273	
5	62			60	10	0,182	0,455	
6	58			61	9	0,164	0,618	
7	61			62	8	0,145	0,764	
8	63			63	7	0,127	0,891	
9	60			64	4	0,073	0,964	
10	60			65	2	0,036	1,000	
11	61	Всего наблюдений			55	1		
12	65							
13	62							
14	62							

Рисунок 2.5 – Результат вычислений относительных и накопленных частот для примера 2.3

6. Построим полигон частот по данным в столбце «Абсолютные частоты», как показано на рисунке 2.6 (используем диаграмму типа «точечная, соединенная отрезками»).

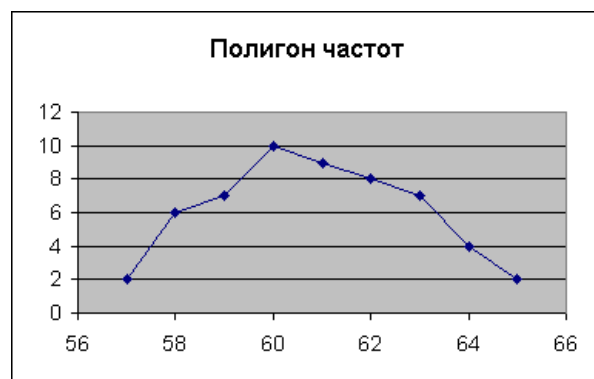



Рисунок 2.6 – Полигон частот для примера 2.3

7. Построим также совместную диаграмму относительных и накопленных частот. Для этого выделим диапазон G1:H10 и вызовем *Мастер диаграмм* кнопкой  на панели инструментов. На вкладке *Нестандартные* выберем тип *График/Гистограмма 2*. После нажатия кнопки *Далее* откроем вкладку *Ряд* и введем в поле *Подписи оси X* диапазон E2:E10. Нажав кнопку *Далее*, введем заголовки осей *X* и *Y*: в поле *Ось X (категорий)* – «Вес», в поле *Ось Y (значений)* – «Относительная частота», в поле *Вторая ось Y (значений)* – «Накопленная частота». Нажмем кнопку *Готово*. После минимального редактирования диаграмма будет иметь такой вид, как показано на рисунке 2.7.

Вспомним, что гистограмма относительных частот есть аналог функции плотности распределения, а график накопленных частот проходит через левые концы «ступенек» эмпирической функции распределения и называется кумулятивной кривой.

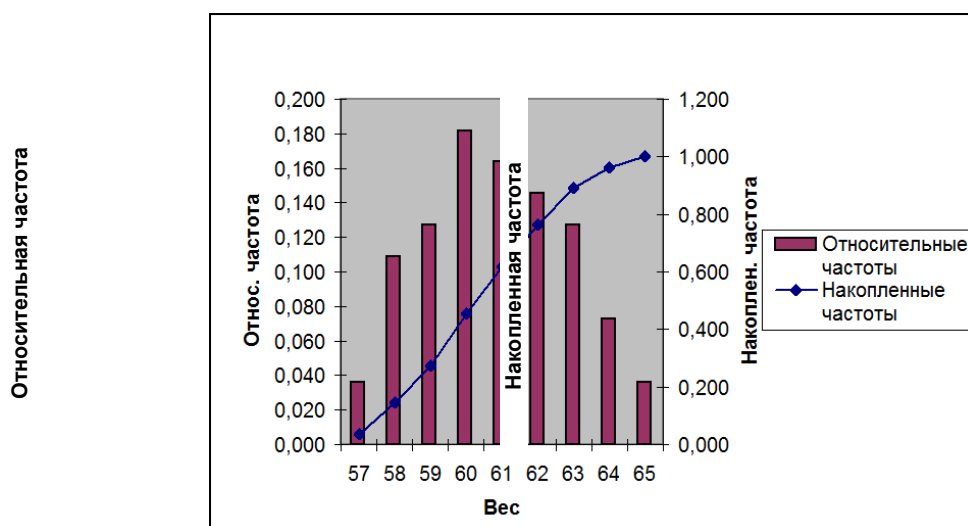


Рисунок 2.7 – Диаграмма относительных и накопленных частот

8. Для построения вариационного ряда можно также использовать процедуру *Гистограмма* надстройки *Пакет анализа*.

Скопируем на чистый лист данные наблюдений из столбца A, а также диапазон значений вариант из столбца E (например, в столбец C).

Зададим команду *Сервис/Анализ данных* и выберем инструмент анализа *Гистограмма*. Заполним окно *Гистограмма*, как показано на рисунке 2.8.

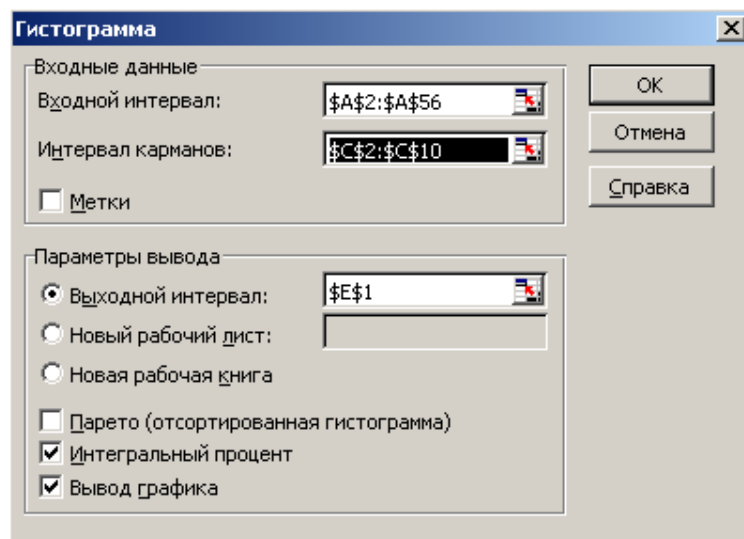


Рисунок 2.8 – Диалоговое окно для построения гистограммы

В поле *Входной интервал* укажем диапазон исследуемых данных наблюдений (A2:A56).

В поле *Интервал карманов* зададим диапазон граничных значений, определяющих выбранные интервалы (карманы). В нашем случае это диапазон возможных значений вариант (C2:C10). Они должны быть введены в возрастающем порядке. Процедура *Гистограмма* вычислит число попаданий данных между началом интервала и соседним большим по порядку. При этом включаются значения на нижней границе интервала и не включаются на верхней. Если не задавать информацию в этом поле, то Excel подберет диапазон карманов автоматически.

Переключатель *Параметры вывода* установим в положение *Выходной интервал* и в соответствующем поле укажем адрес ячейки, в которую будет помещен левый верхний угол результирующей таблицы (E1).

Следует также установить флажки *Вывод графика* и *Интегральный процент* (для дополнительного вывода накопленных частот).

После нажатия кнопки *OK* на рабочем листе Excel появляется таблица и диаграмма (рисунок 2.9). Очевидно, что в столбце «Интегральный процент» показаны накопленные частоты в процентном формате. Диаграмма аналогична рисунку 2.7 (с той разницей, что выводятся не относительные, а абсолютные частоты).

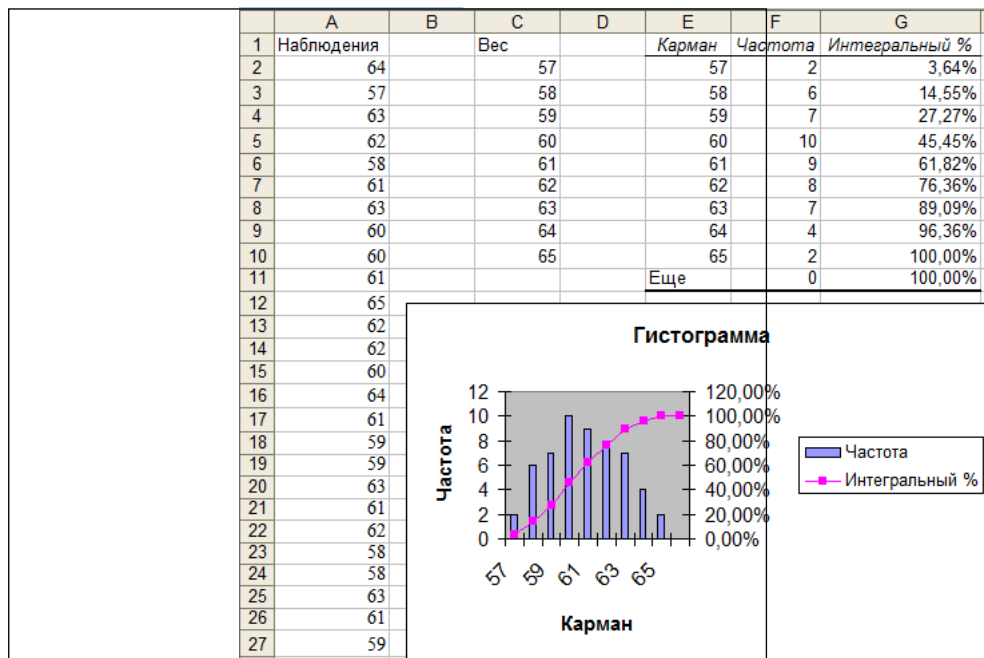


Рисунок 2.9 – Таблица и диаграмма, созданные надстройкой *Пакет анализа*

**Пример 2.4.** Построить интервальный вариационный ряд, используя данные таблицы 2.5.

*Решение*

Разобьем диапазон наблюдавшихся значений [57,65] на интервалы шириной 2 (см. пример 2.3). При этом минимальное значение должно попасть внутрь первого интервала. Данные в столбце «Варианты» (рисунок 2.10) интерпретируются как правые границы интервалов. Значения, которые дает функция *ЧАСТОТА()*, – это

частоты попадания в интервал. При этом, если значение случайной величины попадает на границу интервала, то оно учитывается в левом интервале. Что касается самого первого значения в столбце «Варианты», то для него функция ЧАСТОТА() дает количество наблюдений, меньших или равных ему.

Остальные расчеты полностью аналогичны примеру 2.3. На рисунке 2.10 показан результат вычислений абсолютных и относительных частот для интервального ряда.

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2		64	Минимум	57		8	0,145	0,145
3		57			60	17	0,309	0,455
4		63			62	17	0,309	0,764
5		62			64	11	0,200	0,964
6		58			66	2	0,036	1,000
7		61		Всего наблюдений		55	1	
8		63						
9		60						
10		60						
11		61						

Рисунок 2.10 – Результаты расчетов частот для интервального ряда

Аналогично работает и надстройка *Пакет анализа*. На рисунке 2.11 показан результат работы инструмента *Гистограмма*, которому был задан интервал карманов с шагом 2.

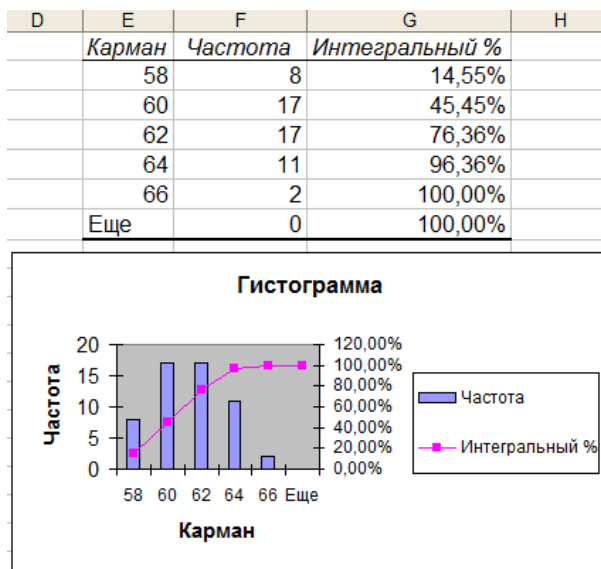


Рисунок 2.11 – Результат работы инструмента *Гистограмма* для примера 2.4

## 2.2. Точечные и интервальные оценки характеристик случайной величины

По выборочным данным часто требуется оценить (приблизительно найти) параметры распределения исследуемой случайной величины. Выборочная оценка некоторого параметра (например, математического ожидания или дисперсии) по сути дела сама является случайной величиной и должна удовлетворять определенным требованиям: быть несмещенной, эффективной и состоятельной.

*Несмещенной* называется оценка, если ее математическое ожидание равно истинному значению характеристики генеральной совокупности.

*Состоятельной* называют статистическую оценку, если с увеличением размера выборки она стремится к значению характеристики генеральной совокупности с вероятностью, близкой к единице.

*Эффективной* называется оценка, которая имеет минимальную дисперсию при заданном объеме выборки  $n$ .

Среди выборочных характеристик случайной величины выделяют показатели, относящиеся к центру распределения (меры положения), показатели рассеяния вариантов (меры рассеяния) и меры формы распределения.

Пусть дана выборка объема  $n$ :  $\{x_1, x_2, \dots, x_n\}$ .

К показателям, характеризующим центр распределения, относят выборочное среднее, а также выборочные моду и медиану.

*Выборочным средним* называется среднее арифметическое выборочных значений:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (2.1)$$

Если эти выборочные значения расположены в виде вариационного ряда  $x_1, x_2, \dots, x_k$  с соответствующими частотами  $n_1, n_2, \dots, n_k$ , то выборочное среднее можно рассчитать по формуле

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i. \quad (2.2)$$

Выборочное среднее является несмещенной оценкой математического ожидания.

*Выборочной модой (Мо)* называется варианта, имеющая наибольшую частоту. Если несколько соседних значений имеют наибольшую частоту, то модой является их среднее арифметическое. Если две или более несмежных вариантов имеют наибольшую частоту, то ряд называется полимодальным (или мультимодальным). В этом случае проверяют, не распадается ли выборка (например, покупатели) на группы (сегменты). Если же все варианты встречаются одинаково часто, то ряд моды не имеет.

*Выборочной медианой (Ме)* называется середина вариационного ряда, т. е. половина вариантов больше медианы, а половина – меньше ее. Если количество членов вариационного ряда нечетное, то медиана – это значение варианты, которое расположено посередине, т. е. элемент с номером  $(n + 1)/2$ . Если число членов ряда четное, то медиана равна среднему элементов с номерами  $[n/2]$  и  $[n/2] + 1$ .

Мода как показатель, характеризующий центр распределения, на практике используется редко. Медиану используют в том случае, если выборка содержит выбросы (очень большие или очень маленькие значения), так как при этом выборочное среднее не отражает типичного значения.

Основными показателями рассеяния вариантов являются интервал, выборочная дисперсия, стандартное отклонение и коэффициент вариации.

*Интервал (размах варьирования)* – это разница между максимальным и минимальным значениями элементов выборки.

*Выборочной дисперсией* называется величина

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2.3)$$

Она характеризует степень разброса элементов выборки относительно среднего значения (оценка дисперсии генеральной совокупности).

Заметим, что в отличие от формулы для расчета дисперсии генеральной совокупности, здесь сумма делится не на  $n$ , а на  $n - 1$ . Можно показать, что в этом случае оценка получается несмещенной. В литературе иногда выборочную дисперсию, рассчитанную по формуле (2.3), называют *исправленной дисперсией*.

*Выборочным средним квадратическим отклонением (стандартным отклонением)* называется величина

$$s = \sqrt{s^2}.$$

Эта величина также отражает степень разброса относительно среднего, но в отличие от дисперсии выражается в тех же единицах, что и значение случайной величины.

*Коэффициентом вариации* называют выраженное в процентах отношение выборочного стандартного отклонения к выборочной средней:

$$V = \frac{s}{\bar{x}} \cdot 100\%.$$

Этот показатель служит для сравнения степени рассеяния двух вариационных рядов: тот из рядов имеет большее рассеяние по отношению к выборочной средней, у которого коэффициент вариации больше. Коэффициент вариации – безразмерная величина, поэтому он пригоден для сравнения вариационных рядов, варианты которых имеют различную размерность (например, варианты одного ряда выражены в сантиметрах, а другого – в граммах).

К мерам формы распределения относятся асимметрия и эксцесс. Эти показатели характеризуют степень отклонения эмпирического (полученного в результате наблюдений или экспериментов) распределения от теоретического нормального распределения. Для нормального распределения асимметрия и эксцесс равны нулю. Выборочные асимметрия ( $A_s$ ) и эксцесс ( $E_k$ ) рассчитываются по формулам:

$$A_s = \frac{n}{(n-1) \cdot (n-2)} \sum_{i=1}^k \left( \frac{x_i - \bar{x}}{s} \right)^3; \quad (2.6)$$

$$E_k = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^k \left( \frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}, \quad (2.7)$$

где  $x_i$  – варианта с частотой  $n_i$  ( $i = 1, 2, \dots, k$ );  
 $n$  – объем выборки;  
 $\bar{x}$  – выборочное среднее;  
 $s$  – выборочное стандартное отклонение.

*Асимметрия* характеризует степень симметрии расположения значений данных относительно среднего. Если у эмпирического распределения длинная и более пологая часть кривой плотности распределения расположена справа от моды, то асимметрия положительна. Если же длинная и пологая часть этой кривой расположена слева от моды, то асимметрия отрицательна (рисунок 2.12).

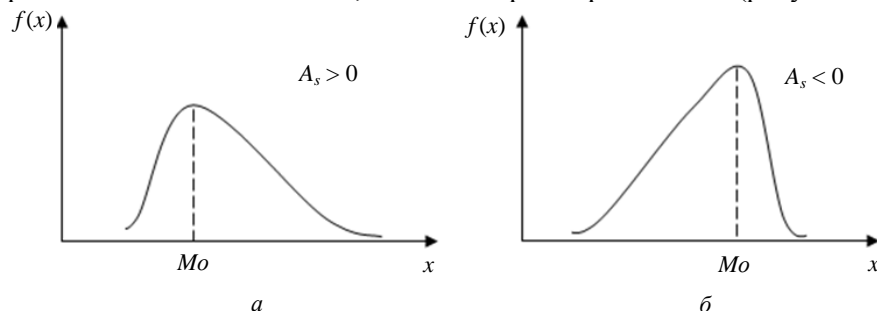


Рисунок 2.12 – Асимметрия, характеризующая форму эмпирической кривой распределения

*Экссесс* характеризует крутизну подъема кривой плотности распределения по сравнению с нормальной кривой: если эксцесс положителен, то кривая имеет более высокую и острую вершину, а если отрицателен – более низкую и пологую вершину (рисунок 2.13).

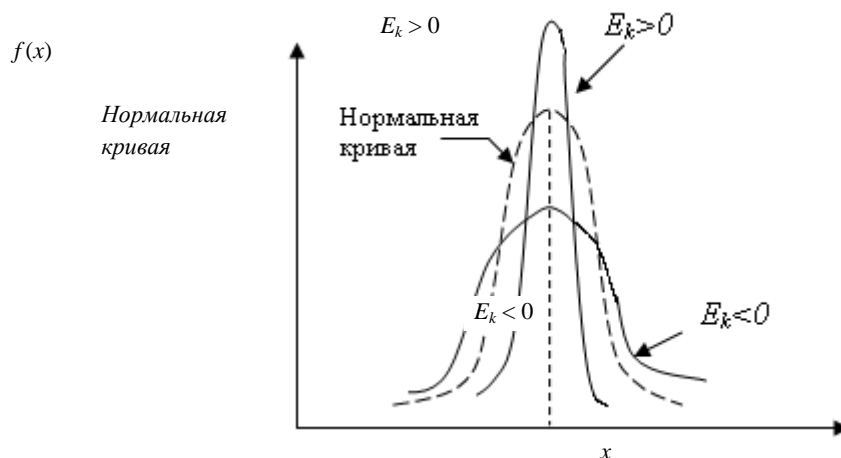


Рисунок 2.13 – Экссесс, характеризующий крутизну эмпирической кривой распределения

В пакете MS Excel для определения выборочных оценок параметров распределения используются следующие функции:

1. СРЗНАЧ() – вычисление среднего арифметического аргументов (т. е. выборочного среднего).
2. МЕДИАНА() – нахождение медианы заданной выборки, т. е. элемента, число значений меньше и больше которого в выборке равны.
3. МОДА() – вычисление наиболее часто встречающегося в выборке значения.
4. ДИСП() – вычисление выборочной дисперсии.
5. СТАНДОТКЛОН() – вычисление выборочного стандартного отклонения.
6. ЭКСЦЕСС() – вычисление оценки эксцесса по выборочным данным.
7. СКОС() – вычисление оценки асимметрии выборочного распределения.

Кроме того, в надстройке *Пакет анализа* имеется инструмент *Описательная статистика*, который позволяет получить все выборочные характеристики случайной величины.

*Точечной* называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, – точечные. Их значения могут значительно отличаться от соответствующих характеристик генеральной со-

вокупности (особенно при малых объемах выборки). Для правильной оценки этого отклонения используется метод доверительных интервалов. Этот метод разработал американский статистик Ю. Нейман, исходя из идей английского статистика Р. Фишера.

*Интервальной* называют оценку, определяемую двумя числами – концами интервала, который с заданной вероятностью «накрывает» значение оцениваемого параметра генеральной совокупности (рисунок 2.14).

Пусть, например, необходимо построить доверительный интервал для математического ожидания  $M(X)$ . Для заранее выбранного уровня значимости  $\alpha$  ( $0 < \alpha < 1$ ) по выборке определяются два числа  $a$  и  $b$ , ( $a < b$ ), между которыми с вероятностью  $p = 1 - \alpha$  находится неизвестный параметр  $M(X)$ :  $P(a < M(X) < b) = 1 - \alpha$ .

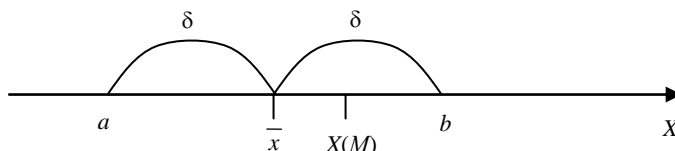


Рисунок 2.14 – Доверительный интервал, «накрывающий» значение параметра распределения генеральной совокупности

Вероятность  $p = 1 - \alpha$  попадания значения параметра распределения генеральной совокупности в доверительный интервал называется *доверительной вероятностью (надежностью)*.

*Уровень значимости  $\alpha$*  – это вероятность противоположного события (непопадания истинного значения параметра в доверительный интервал). В качестве уровня значимости обычно принимают значения  $\alpha = 0,1$  (0,05; 0,01), которые определяют 90%-ные (95-, 99%-ные) доверительные интервалы соответственно.

Концы интервала  $a$  и  $b$  называются *доверительными* (нижней и верхней) *границами*. Эти величины определяются по результатам выборки и, следовательно, являются случайными.

Точечной оценкой математического ожидания является выборочное среднее  $\bar{x}$ . Границы доверительного интервала находятся как  $a = \bar{x} - \delta$ ,  $b = \bar{x} + \delta$ , где  $\delta > 0$  – *точность* доверительного интервала, которая либо задается заранее, либо вычисляется.

Задача определения доверительного интервала является довольно сложной. Ее решение зависит от объема выборки (большой или малый), от того, известен ли закон распределения исследуемой случайной величины, а также известны ли какие-либо параметры распределения. Например, при расчете доверительного интервала для математического ожидания может быть известно (или не известно) значение дисперсии генеральной совокупности (истинное значение дисперсии). Рассмотрим только один, наиболее распространенный на практике способ определения доверительного интервала для математического ожидания.

Пусть известно, что случайная величина, из которой получена выборка объема  $n$ , имеет нормальный закон распределения. Истинное значение дисперсии этой случайной величины будем считать неизвестным. Рассмотрим два случая:

1. Если число наблюдений достаточно велико ( $n > 30$ ), то с вероятностью 0,95 значение математического ожидания попадает в доверительный интервал  $\bar{x} \pm 2m$ , а с вероятностью 0,99 – в доверительный интервал  $\bar{x} \pm 3m$ , где величина  $m$ , определяемая по формуле

$$m = \frac{s}{\sqrt{n}},$$

называется *стандартной ошибкой*, или *ошибкой среднего*.

2. Если число наблюдений мало ( $n \leq 30$ ), то доверительный интервал определяется по формулам:

$$m = \bar{x} \pm \delta; \quad (2.9)$$

$$\delta = t_{\alpha, n-1} \frac{s}{\sqrt{n}},$$

где  $t_{\alpha, n-1}$  – критическая точка распределения Стьюдента с числом степеней свободы  $n - 1$  и уровнем значимости  $\alpha$ .

Для вычисления критической точки распределения Стьюдента в MS Excel можно воспользоваться функцией СТЬЮДРАСПОБР(). Например, если было 20 наблюдений и нужно получить доверительный



интервал, в который математическое ожидание попадает с вероятностью 0,95 ( $\alpha = 1 - 0,95 = 0,05$ ), то аргументы этой функции будут следующие: СТЮДРАСПОБР(0,05;19).

Аналогично доверительному интервалу для математического ожидания можно определять доверительные интервалы для других выборочных характеристик.

Иногда полученная точность не удовлетворяет пользователя, так как дает слишком широкий диапазон, в который попадает математическое ожидание с вероятностью  $p$ . Чем меньше точность доверительного интервала, тем ближе выборочная оценка к соответствующему генеральному показателю.

Точность зависит от числа наблюдений. Можно определить число наблюдений, которое необходимо для достижения заданной точности  $\delta$  по следующей формуле:

$$n(\delta) \geq t_{\alpha, n-1} \frac{s^2}{\delta^2} + 1.$$

**Пример 2.5.** По данным о весе учащихся из таблицы 2.5 (см. пример 2.3) с помощью MS Excel рассчитать выборочные характеристики распределения. Найти 95%-ный доверительный интервал для математического ожидания. Определить, сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,2.

*Решение*

Введем эмпирические данные о весе учащихся на чистый лист Excel и оформим его, как показано на рисунке 2.15 (можно скопировать данные с листа, который использовался при решении предыдущего примера).

	A	B	C	D	E	F	G	H
1	Наблюдения		Выборочные оценки (используя стандартные функции)				Выборочные оценки (пакет анализа)	
2	64	Среднее	60,855				Наблюдения	
3	57	Дисперсия	4,201					
4	63	стандартное отклонение	2,050				Среднее	60,85455
5	62	мода	60				Стандартная ошибка	0,276362
6	58	медиана	61				Медиана	61
7	61	эксцесс	-0,744				Мода	60
8	63	асимметрия	0,096				Стандартное отклонение	2,049554
9	60	количество наблюдений	55				Дисперсия выборки	4,200673
10	60	Расчет доверительного интервала					Эксцесс	-0,7437
11	61	большое число наблюдений					Асимметричность	0,096321
12	65	точность	0,5527				Интервал	8
13	62	нижняя граница	60,3018				Минимум	57
14	62	верхняя граница	61,4073				Максимум	65
15	60	малое число наблюдений					Сумма	3347
16	64						Счет	55
17	61	точность	0,5541				Уровень надежности(95,0%)	0,554072
18	59	нижняя граница	60,3005					
19	59	верхняя граница	61,4086					
20	63							
21	61	Количество наблюдений $n(0,2)$ для точности				0,2		
22	62	211,5461						

Рисунок 2.15 – Лист Excel с расчетом выборочных характеристик распределения

В ячейки D2:D9 введем стандартные функции Excel категории *Статистические* СРЗНАЧ(), ДИСП(), СТАНДОТКЛОН(), МОДА(), МЕДИАНА(), ЭКСЦЕСС(), СКОС() и СЧЕТ(). Аргументом всех этих функций является диапазон выборочных значений веса A2:A56.

Отметим, что распределение веса студентов является достаточно симметричным (асимметрия равна 0,096 и близка к нулю), а эксцесс имеет небольшое отрицательное значение (–0,744). Это означает, что распределение веса имеет более низкую и пологую вершину по сравнению с нормальным распределением.

Аналогичные данные можно получить с помощью инструмента *Описательная статистика* надстройки *Пакет анализа*. Зададим команду *Сервис/Анализ данных* и выберем инструмент *Описательная статистика*. Заполним диалоговое окно, как показано на рисунке 2.16.

Данные каждой выборки должны быть расположены в одном столбце или одной строке (можно получить информацию сразу по нескольким выборкам). Переключатель *Группирование* в нашем случае установлен в положение *По столбцам*, так как эмпирические данные занесены в первый столбец.

Флажок *Метки в первой строке* установлен, так как входной интервал включает и заголовок данных в первой строке (слово «Наблюдения»).

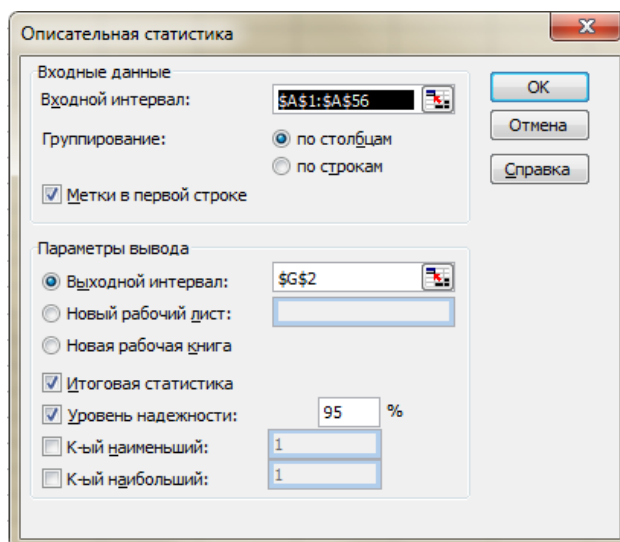


Рисунок 2.16 – Диалоговое окно для вывода описательной статистики

Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, так как нужно получить результаты расчетов на текущем листе Excel. В соответствующем поле указан адрес левой верхней ячейки выходного диапазона (G2).

Установим флажок *Итоговая статистика* для вывода на листе Excel выборочных характеристик и флажок *Уровень надежности* для вывода точности доверительного интервала с заданным уровнем надежности ( $p = 0,95$ ).

После заполнения этого диалогового окна нажмем кнопку *OK*. Результаты расчетов с помощью *Пакета анализа* показаны на рисунке 2.15 в столбцах G и H.

Все полученные характеристики были рассмотрены ранее, за исключением четырех последних:

- минимум – значение минимального элемента выборки;
- максимум – значение максимального элемента выборки;
- сумма – сумма значений элементов выборки;
- счет – количество элементов в выборке.

При расчете доверительного интервала для математического ожидания необходимо учесть, что число наблюдений в данном примере достаточно велико ( $55 > 30$ ), поэтому можно считать, что интервал  $\bar{x} \pm 2m$  покрывает математическое ожидание генеральной совокупности с вероятностью 0,95. Значение стандартной ошибки рассчитано *Пакетом анализа* в ячейке H5 ( $m = 0,276$ ). Поэтому точность доверительного интервала будет равна:  $\delta = 2m = 2 \cdot 0,276 = 0,553$ . Это значение рассчитано в ячейке D12 по формуле  $=2*H5$ .

В ячейках D13 и D14 найдены также нижняя и верхняя границы доверительного интервала.

В случае, когда количество наблюдений недостаточно ( $n < 30$ ), для расчета доверительного интервала для математического ожидания необходимо использовать формулу (2.10). Если применить ее к данному примеру, то учитывая, что число наблюдений ( $n$ ) равно 55 (значение в ячейке D9), а требуемая надежность ( $p$ ) составляет 0,95 (уровень значимости  $\alpha = 1 - p = 0,05$ ), расчет по формуле (2.10) в ячейке D17 должен был иметь вид

$$=СТЫЮДРАСПОБР(0,05;D9-1)*D4/КОРЕНЬ(D9).$$

Результат расчета по этой формуле совпадает со значением в ячейке H17, которое дает *Пакет анализа*. Таким образом, *Пакет анализа* всегда рассчитывает точность доверительного интервала по формуле (2.10), ориентированной на малое количество наблюдений. При этом доверительный интервал получается немного больше (отличие точности – в третьем знаке).

Для полученной точности также рассчитаем нижнюю и верхнюю границы доверительного интервала в ячейках D18 и D19.

Полученная точность доверительного интервала ( $\delta \approx 0,55$ ) превышает заданное значение 0,2. Определим по формуле (2.11), сколько необходимо иметь данных наблюдений для достижения этой точности. Для этого введем, например в ячейку C22, следующую формулу:

$$=СТЫЮДРАСПОБР(0,05;D9-1)*D4^2/0,2^2+1.$$

Полученное значение показывает, что нужно иметь не менее 212 наблюдений.

### 2.3. Общие сведения о проверке статистических гипотез

*Статистической гипотезой* называют утверждение о виде неизвестного распределения случайной величины  $X$  или о параметрах известных распределений, которое может быть проверено по выборочным данным. При этом, если речь идет о параметрах распределения, то гипотеза называется *параметрической*. Если же утверждение касается вида распределения, то гипотеза называется *непараметрической*.

Например, статистическими являются гипотезы:

- генеральная совокупность имеет нормальный закон распределения;
- дисперсии двух нормально распределенных случайных величин равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения (непараметрическая гипотеза). Во второй гипотезе утверждение касается параметров двух известных распределений (параметрическая гипотеза). Гипотеза «на Марсе есть жизнь» не является статистической, так как в ней не идет речь ни о виде, ни о параметрах распределения.

Обычно наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. Основную выдвинутую гипотезу называют *нулевой* и обозначают  $H_0$ . Противоречащую ей гипотезу называют *альтернативной* (*конкурирующей*) и обозначают  $H_1$ .

Например, если нулевая гипотеза состоит в предположении, что закон распределения генеральной совокупности является нормальным, то альтернативная – в том, что он нормальным не является. При этом не указывается другой вид закона распределения, а только отрицается его нормальность.

Выдвинутая гипотеза может быть правильной или неправильной, но судить об этом мы можем только на основании выборки из генеральной совокупности. Поскольку выборка не является полной информацией, может быть принято неверное решение при выборе между основной и альтернативной гипотезами. Причем ошибки, возникающие при проверке гипотезы, могут быть двух родов.

*Ошибка первого рода* состоит в том, что будет отвергнута правильная нулевая гипотеза.

*Ошибка второго рода* состоит в том, что будет принята неправильная нулевая гипотеза.

Ситуации, которые возможны при проверке гипотезы, представлены в таблице 2.6.

Таблица 2.6 – Ошибки первого и второго рода

Результаты проверки гипотезы	Возможные состояния гипотезы	
	$H_0$ верна	$H_0$ неверна
Гипотеза $H_0$ отвергается	Ошибка первого рода	Правильный вывод
Гипотеза $H_0$ принимается	Правильный вывод	Ошибка второго рода

Последствия этих ошибок могут быть весьма различными в зависимости от содержания гипотезы.

Чтобы лучше понять определения ошибок, проиллюстрируем эти понятия с помощью аналогии. В больнице врач принимает решение: направлять пациента на операцию или не отправлять. Его проблему можно переформулировать так: ему нужно выбрать между основной гипотезой, что операция необходима, и альтернативной гипотезой, утверждающей, что операция не нужна. При этом врач может ошибиться. Допустим, операция нужна, но она не делается (основная гипотеза верна, но она отвергается). В этом случае врач делает ошибку первого рода. Если операция не нужна, но она делается, т. е. принимается основная гипотеза, когда она неверна, то врач делает ошибку второго рода.

В приведенном примере последствия ошибки первого рода более серьезны, так как отсутствие нужной операции может привести к смертельному исходу. «Лишняя» операция – это очень неприятно, но не так опасно. Можно привести и другие примеры, когда последствия ошибки второго рода более существенны, чем ошибки первого рода.

Кроме того, следует отметить, что при проверке статистических гипотез стараются быть сдержанными в выражениях. Фраза «гипотеза  $H_0$  принимается» на самом деле означает не то, что она справедлива, а то, что выборочные данные не противоречат гипотезе.

Для проверки нулевой гипотезы используют специально подобранную случайную величину, закон распределения которой известен. Эту величину обозначают через  $U$  или  $Z$ , если она распределена нормально,  $F$  – по закону Фишера,  $T$  – по закону Стьюдента,  $\chi^2$  – по закону «хи-квадрат». В целях общности обозначим в данном разделе эту величину через  $K$ .

*Статистическим критерием* называют случайную величину ( $K$ ), которая служит для проверки нулевой гипотезы.

После выбора критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значение критерия, при которых нулевая гипотеза отвергается, а другое – при которых она принимается.

*Критической областью* называется совокупность значений критерия, при которых нулевую гипотезу отвергают.

*Областью принятия гипотезы* (областью допустимых значений) называют совокупность значений критерия, при которых гипотезу принимают.

Критическими точками ( $k_{кр}$ ) называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую неравенством  $K > k_{кр}$ , где  $k_{кр}$  – положительное число (рисунок 2.17, а).

Левосторонней называют критическую область, определяемую неравенством  $K < k_{кр}$ , где  $k_{кр}$  – отрицательное число (рисунок 2.17, б).

Двусторонней называют критическую область, определяемую неравенствами  $K < k_1$ ,  $K > k_2$ , где  $k_2 > k_1$ . В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенством  $|K| > k_{ед}$  (рисунок 2.17, в).

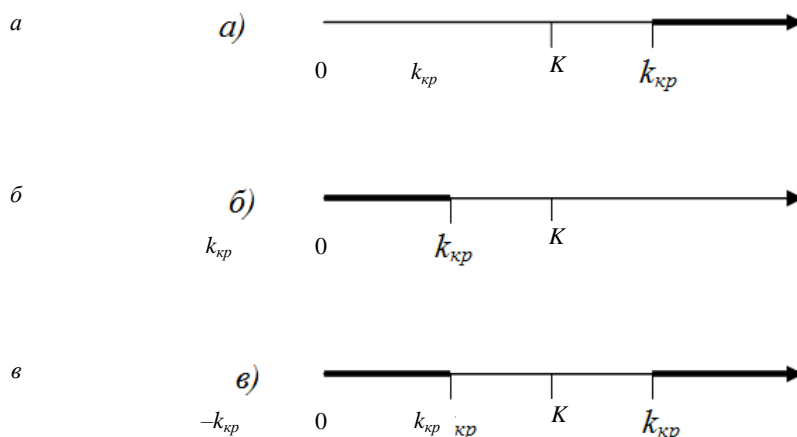


Рисунок 2.17 – Виды критической области

Для определения критической области задают достаточно малую вероятность – уровень значимости  $\alpha$ . В качестве уровня значимости чаще всего выбирают одно из чисел 0,1; 0,05; 0,01. Критические точки определяются исходя из требования, чтобы при условии справедливости нулевой гипотезы вероятность того, что критерий  $K$  попадет в критическую область, была равна принятому уровню значимости. Например, для правосторонней критической области  $k_{кр}$  выбирается из условия  $P(K > k_{ед}) = \alpha$ .

Суть проверки статистической гипотезы состоит в том, что по данным выборки вычисляют частное (наблюдаемое) значение критерия ( $K_{набл}$ ), а затем проверяют, попадает ли оно в критическую область. В частности, для правосторонней критической области проверяют условие  $K_{набл} > k_{кр}$ . Если это условие выполняется, то нулевую гипотезу отвергают. Если же наблюдаемое значение попадает в область принятия гипотезы, то нет оснований отвергнуть нулевую гипотезу.

Отметим, что наблюдаемое значение критерия может попасть в критическую область не потому, что нулевая гипотеза ложна, а по другим причинам (малый объем выборки, недостатки методики эксперимента и др.) В этом случае, отвергнув правильную нулевую гипотезу, совершают ошибку первого рода. Вероятность этой ошибки равна уровню значимости  $\alpha$ .

Поскольку уровень значимости сознательно задается очень малым, то обеспечивается малая вероятность ошибки первого рода. Значительно труднее добиться того, чтобы вероятность ошибки второго рода была малой. Как правило, ее можно уменьшить, если увеличить число анализируемых наблюдений. Поэтому так необходимы большие выборки. Если выборка маленькая ( $n < 30$ ), то проверить гипотезу можно, но минусом будет большая вероятность ошибки второго рода. Профессиональные статистики в таких случаях часто увеличивают уровень значимости (например, до 0,15 или 0,2), чтобы сделать вероятности ошибок первого и второго рода сопоставимыми.

Обозначим вероятность ошибки второго рода  $\beta$  (вероятность принять гипотезу  $H_0$  при условии, что она неверна). Тогда вероятность отвергнуть гипотезу  $H_0$  при условии, что она неверна (т. е. вероятность принять верное решение), равна  $1 - \beta$  и называется *мощностью критерия*. Вид статистического критерия и способ построения критической области выбираются таким образом, чтобы получить как можно большую мощность.

Таким образом, стандартный алгоритм проверки статистических гипотез следующий:

1. Выдвигается нулевая гипотеза  $H_0$  и альтернативная ей гипотеза  $H_1$ .
2. Выбирается уровень значимости  $\alpha$  (обычно  $\alpha = 0,1; 0,05; 0,01$ ).
3. Выбирается статистический критерий ( $K$ ), имеющий известный закон распределения.
4. На основании выбранного уровня значимости определяется критическая область критерия.
5. Вычисляется наблюдаемое значение критерия ( $K_{набл}$ ) по выборочным данным.

6. Если наблюдаемое значение критерия попадает в критическую область, то нулевая гипотеза отвергается, в противном случае  $H_0$  принимается (говорят, что выборочные данные не противоречат гипотезе).

Наряду с критерием значимости для проверки нулевой гипотезы существует правило, основанное на  $p$ -значении.

Чтобы понять смысл этого показателя, рассмотрим пример для правосторонней критической области (рисунок 2.18). Очевидно, что чем больше заданный уровень значимости (больше вероятность ошибки первого рода, которую мы допускаем), тем шире критическая область, т. е. критическая точка расположена левее. На рисунке 2.18 условно показано расположение критических точек для двух уровней значимости ( $\alpha = 0,01$  и  $\alpha = 0,05$ ). Если наблюдаемое значение критерия примет показанное на рисунке значение, то при уровне значимости  $\alpha = 0,05$  гипотеза  $H_0$  будет отвергнута, а при  $\alpha = 0,01$  принята.

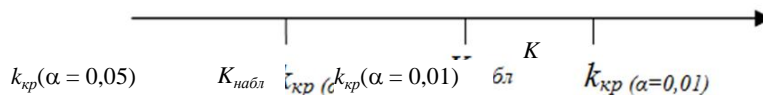


Рисунок 2.18 – Критическая область при разных уровнях значимости

Возникает вопрос: каков тот пограничный уровень значимости, при котором  $K_{набл}$  начинает попадать в критическую область, что ведет к отказу от нулевой гипотезы (при условии ее справедливости). Эта величина и называется  $p$ -значением.

$P$ -значение представляет собой вероятность того, что какое-либо значение статистического критерия будет больше наблюдаемого по данной выборке значения  $p = P(K \geq K_{набл})$  при условии справедливости нулевой гипотезы.

Таким образом,  $p$ -значение есть вероятность ошибки первого рода для ситуации, когда наблюдаемое значение критерия было бы критической точкой.

Очевидно, что  $p = P(K \geq K_{набл}) = 1 - P(K < K_{набл}) = 1 - F(K_{набл})$ . Поэтому можно интерпретировать  $p$ -значение как величину отрезка от значения интегральной функции распределения в точке  $K_{набл}$  до единицы (рисунок 2.19). Соответственно,  $\alpha = P(K \geq K_{кр}) = 1 - P(K < K_{кр}) = 1 - F(K_{кр})$ . Поскольку интегральная функция распределения является монотонно возрастающей, то если  $K_{кр} < K_{набл}$  (нулевая гипотеза должна быть отвергнута), то  $F(K_{кр}) < F(K_{набл})$  и, следовательно,  $\alpha > p$ .

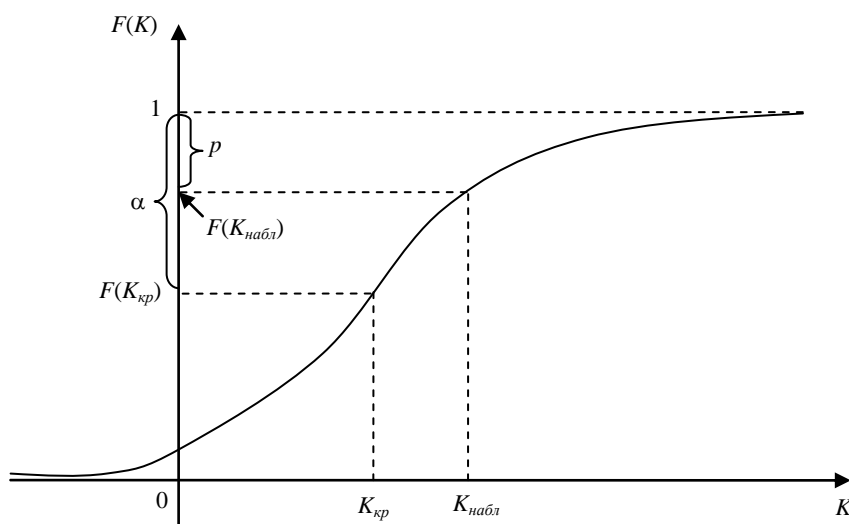


Рисунок 2.19 – Графическая иллюстрация понятия  $p$ -значения

Верно и обратное: если  $\alpha > p$ , то наблюдаемое значение критерия оказывается правее критической точки, что означает необходимость отвергнуть нулевую гипотезу. В связи с этим можно сформулировать *правило проверки статистической гипотезы*, основанное на использовании  $p$ -значения:

Найденное  $p$ -значение сравнивается с уровнем значимости  $\alpha$ . Если  $p < \alpha$ , то гипотеза  $H_0$  отклоняется. Если  $p > \alpha$ , то нет оснований отвергнуть гипотезу  $H_0$ .

И в целом, чем больше  $p$ -значение, тем больше оснований принять нулевую гипотезу, поскольку для любого уровня значимости, большего  $p$ , гипотеза будет принята.

## 2.4. Оценка соответствия выборочных данных теоретическому закону распределения

Важной задачей при анализе одной выборки является оценка соответствия изучаемой случайной величины какому-либо теоретическому закону распределения. По выборочным данным наиболее часто проверяется предположение о нормальном распределении генеральной совокупности. Для оценки соответствия экспериментальных данных теоретическому закону распределения используются следующие методы:

- графический метод;
- метод оценки выборочных параметров формы распределения (асимметрия и эксцесс);
- проверка критериев согласия.

Графический метод состоит в визуальном сравнении гистограммы выборочного распределения и графика теоретической функции плотности нормального распределения. На рисунках 2.20–2.22 приведены примеры гистограмм, по которым можно предположить, что генеральная совокупность подчиняется тому или иному закону распределения. Например, если гистограмма близка по форме к симметричному колоколу, то предполагается нормальный закон распределения.

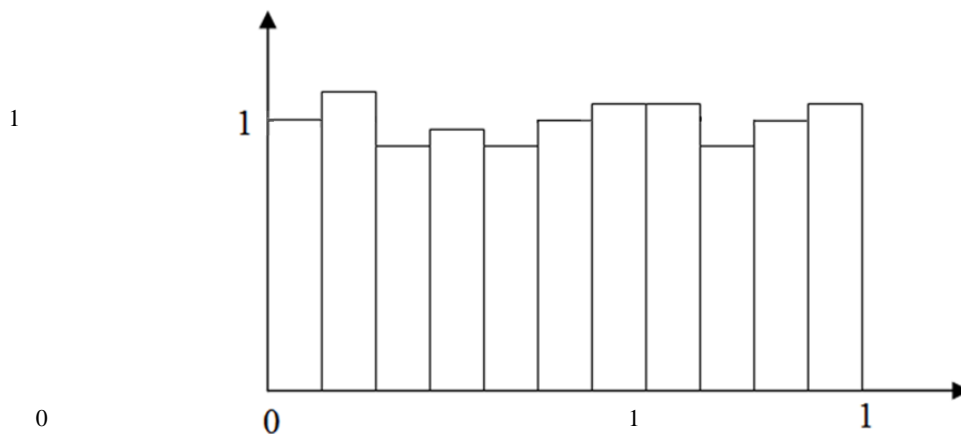


Рисунок 2.20 – Гистограмма случайной величины, имеющей равномерное распределение

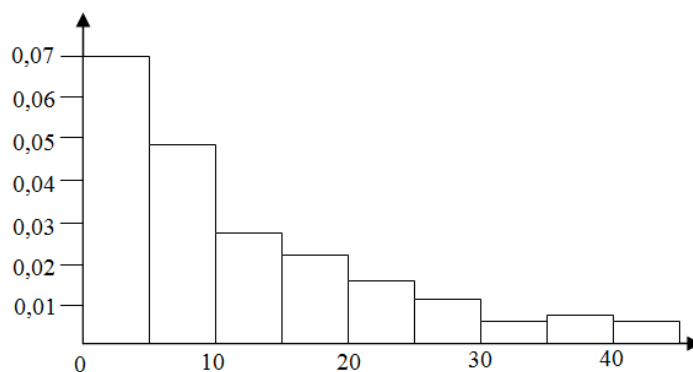


Рисунок 2.21 – Гистограмма случайной величины, имеющей экспоненциальное распределение

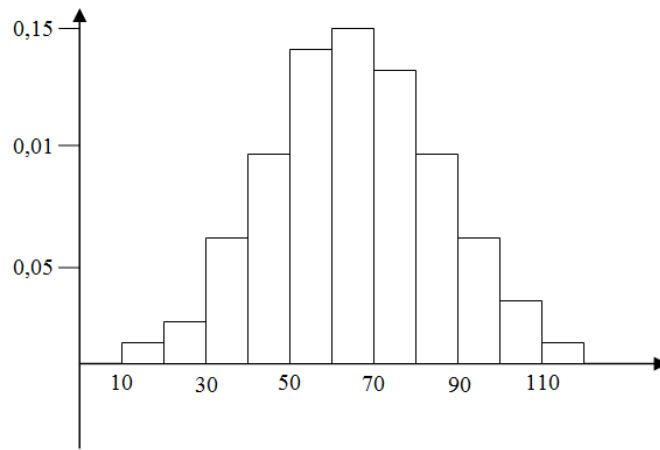


Рисунок 2.22 – Гистограмма случайной величины, имеющей нормальный закон распределения

Для оценки соответствия эмпирических данных нормальному закону распределения можно использовать асимметрию и эксцесс. Как уже отмечалось, для нормального распределения эти показатели равны нулю. Если соответствующие выборочные значения лежат в диапазоне от  $-0,2$  до  $0,2$ , то считается, что предположение о нормальности распределения не противоречит исходным данным. К сожалению, практика показала, что оба показателя неустойчивы, переменчивы. Кроме того, должно быть достаточно большое число наблюдений ( $n > 100$ ), что не всегда возможно обеспечить. Поэтому в настоящее время асимметрию и эксцесс используют редко.

Наиболее убедительные результаты дает использование методов проверки статистических гипотез. Статистические критерии, которые используются для проверки гипотезы о законе распределения, называются *критериями согласия*. Наиболее часто применяются критерий Пирсона и критерий Колмогорова-Смирнова. В некоторых специализированных статистических пакетах реализован также критерий Шапиро-Уилка.

**Критерий Пирсона (критерий  $\chi^2$ ).** Для применения критерия Пирсона желательно, чтобы объем выборки ( $n \geq 40$ ), выборочные данные были бы сгруппированы в интервальный ряд, а в каждом интервале находилось не менее 5 наблюдений. Если это условие не выполняется, то интервалы могут объединяться с соседними.

Суть критерия Пирсона состоит в сравнении эмпирических (наблюдаемых) частот и теоретических частот (т. е. вычисленных в предположении некоторого закона распределения). Пусть для определенности проверяется соответствие данных нормальному закону распределения.

Выдвигается гипотеза  $H_0$ : распределение генеральной совокупности, из которой взята выборка, не отличается от нормального. Альтернативная гипотеза –  $H_1$ : распределение генеральной совокупности отличается от нормального.

Алгоритм *первого способа* проверки гипотезы состоит в следующем:

1. Рассчитываются теоретические частоты для интервалов группировки по формуле нормального распределения. Вероятность попадания в интервал  $P(x_{i-1} \leq X \leq x_i) = F(x_i) - F(x_{i-1})$  (см. раздел 1). Чтобы получить теоретическую частоту, нужно эту вероятность умножить на число наблюдений:

$$n'_i = n(F(x_i) - F(x_{i-1})),$$

где  $F(x)$  – интегральная функция нормального распределения;

$x_{i-1}$  и  $x_i$  – концы интервала группировки;

$n$  – объем выборки.

Теоретически при расчете функции распределения параметры предполагаемого распределения должны быть известны. Например, для нормального распределения должны быть известны  $\mu$  и  $\sigma$ . Однако на практике чаще всего истинные значения параметров заменяются их выборочными оценками ( $\bar{x}$  и  $s$ ).

2. В качестве статистического критерия используется случайная величина

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i},$$

где  $n_i$  – фактическая (наблюдаемая) частота;

$n'_i$  – теоретическая частота попадания значения случайной величины в  $i$ -й интервал;

$k$  – число интервалов группировки.



Этот критерий имеет распределение  $\chi^2$  с  $(k-r-1)$  степенями свободы, где  $r$  – число параметров предполагаемого распределения, которые оцениваются по выборке. Например, для нормального распределения нужно задать два параметра ( $a$  и  $\sigma$ ), поэтому  $r = 2$ . Для распределения Пуассона нужен один параметр  $\lambda$ , поэтому  $r = 1$ .

Очевидно, что чем ближе наблюдаемые частоты к теоретическим, тем меньше будет значение критерия Пирсона (2.13). Поэтому гипотезу  $H_0$  следует отвергнуть при большом значении этого критерия, т. е. можно использовать правостороннюю критическую область. Для выбранного уровня значимости  $\alpha$  с учетом количества степеней свободы  $(k-r-1)$  находится критическая точка  $\chi_{\alpha, k-r-1}^2$ . Для определения этого значения в приложении MS Excel используется функция ХИ2ОБР().

3. Рассчитывается наблюдаемое значение критерия ( $\chi_{\text{наб}}^2$ ), для которого проверяется, попадает ли оно в критическую область. Если  $\chi_{\text{наб}}^2 < \chi_{\alpha, k-r-1}^2$ , то нет оснований отвергнуть гипотезу  $H_0$  о том, что закон распределения является нормальным. Если же  $\chi_{\text{наб}}^2 \geq \chi_{\alpha, k-r-1}^2$ , то гипотеза  $H_0$  отвергается.

*Второй способ* проверки гипотезы с помощью критерия Пирсона состоит в том, что необходимо рассчитать  $p$ -значение. Для этого используется функция ХИ2ТЕСТ(). Результат расчета – предельный уровень значимости, при котором данная выборка дает основание отвергнуть нулевую гипотезу. Если  $p$ -значение меньше уровня значимости  $\alpha$ , то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют нормальному закону распределения. Если же вычисленная вероятность близка к единице, то можно говорить о высокой степени соответствия эмпирических данных нормальному закону.

Основным недостатком критерия Пирсона является то, что вывод зависит от того, как сгруппированы данные. Не существует четкой инструкции, как сделать так, чтобы получить высокую мощность критерия. Говорят, что критерий Пирсона достоверен асимптотически, т. е. чем больше число наблюдений ( $n$ ), тем более правильные результаты.

**Пример 2.6.** На основании данных о весе студентов из таблицы 2.5 (см. пример 2.3) с помощью критерия Пирсона проверить гипотезу о нормальном законе распределения исследуемой случайной величины.

*Решение*

Скопируем данные наблюдений веса на чистый лист Excel и построим вариационный ряд аналогично тому, как это делалось в примере 2.3 (можно использовать функцию ЧАСТОТА() или инструмент *Гистограмма* из *Пакета анализа*). Абсолютные частоты вариант расположим в столбце *Фактические частоты*, как показано на рисунке 2.23.

	A	B	C	D	E	F	G	H	I	J
1	Наблюдения		максимум	65		Варианты	Фактические частоты	Левые границы интервалов	Правые границы интервалов	Теоретические частоты
2	64		минимум	57		57	2	56,5	57,5	1,87
3	57		среднее	60,855		58	6	57,5	58,5	4,10
4	63		стандартное отклон.	2,050		59	7	58,5	59,5	7,10
5	62		асимметрия	0,096		60	10	59,5	60,5	9,73
6	58		эксцесс	-0,744		61	9	60,5	61,5	10,57
7	61					62	8	61,5	62,5	9,10
8	63					63	7	62,5	63,5	6,20
9	60					64	4	63,5	64,5	3,34
10	60					65	2	64,5	65,5	1,43
11	61					Всего наблюдений	55			

Рисунок 2.23 – Расчет фактических и теоретических частот

Рассчитаем выборочные среднее, стандартное отклонение, асимметрию и эксцесс, используя функции СРЗНАЧ(), СТАНДОТКЛОН(), СКОС() и ЭКСЦЕСС().

В ячейке G11 найдем общее число наблюдений, просуммировав фактические частоты.

Чтобы рассчитать теоретические частоты, перейдем к интервальному вариационному ряду. Зададим интервалы группировки так, чтобы наблюдаемые значения вариант стали их серединами. Для этого в ячейке H2 (столбец *Левые границы интервалов*) введем формулу =F2–0,5. Скопируем ее вниз по столбцу. В ячейке I2 (столбец *Правые границы интервалов*) введем формулу =F2+0,5 и скопируем аналогично. Отметим, что если ряд сразу построен как интервальный (см. пример 2.4), то эти действия не нужны.

В столбце *Теоретические частоты* рассчитаем теоретические частоты для нормального распределения. Математическое ожидание и среднеквадратическое отклонение будем считать равными соответствующим выборочным характеристикам. Для этого в ячейку J2 введем формулу, соответствующую выражению (2.12):

$$=(\text{НОРМРАСП}(I2;\$D\$3;\$D\$4;1)-\text{НОРМРАСП}(H2;\$D\$3;\$D\$4;1))*\$G\$11.$$



Функция НОРМРАСП() дает значение интегральной функции нормального распределения для заданного значения  $X$  (первый аргумент I2), при этом математическое ожидание равно значению, записанному в ячейке D3, а стандартное отклонение равно значению в ячейке D4. Последний аргумент этой функции (1) означает, что результатом является значение интегральной функции распределения, а не функции плотности распределения (которое получается при значении этого параметра, равном нулю). Разность значений функции распределения на концах интервала равна вероятности попадания в данный интервал. А для получения абсолютной теоретической частоты это значение нужно еще умножить на число наблюдений в ячейке G11.

Данную формулу нужно скопировать методом автозаполнения в ячейки диапазона J3:J10.

По данным в ячейках G2:G10 и J2:J10 можно построить нестандартную диаграмму типа *График/Гистограмма* 2, как это было описано в примере 2.3 (рисунок 2.24).

Визуальное сравнение гистограммы и графика теоретических частот дает основание полагать, что исследуемое распределение близко нормальному. График исследуемого распределения является симметричным (асимметрия мала), но более пологим, чем график теоретического нормального распределения (эксцесс имеет небольшое отрицательное значение) (рисунок 2.23).

Для оценки степени соответствия исследуемого распределения и теоретического нормального распределения применим критерий согласия  $\chi^2$  (критерий Пирсона). Чтобы выполнить все условия его применения, объединим два первых и два последних интервала. Тогда фактические частоты всех интервалов станут достаточно большими (не менее 5). Объединенные интервалы, а также соответствующие фактические и теоретические частоты показаны на рисунке 2.25. Их значения получаются путем копирования из столбцов G и J (кроме первого и последнего значения, которые рассчитываются путем суммирования).

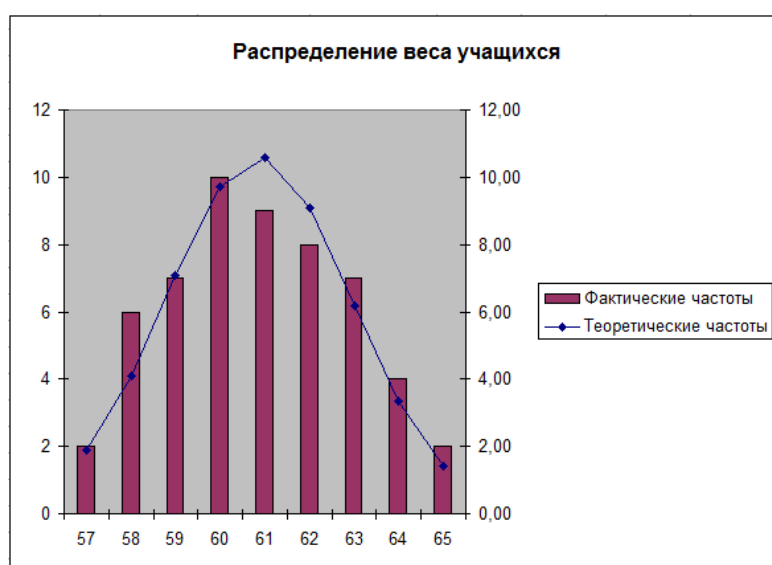


Рисунок 2.24 – Диаграмма фактических и теоретических частот

	F	G	H	I	J
	Объединенные интервалы	Фактические частоты ( $n_i$ )	Теоретические частоты ( $n_i'$ )	$(n_i - n_i')^2 / n_i'$	
13					
14	56,5 - 58,5	8	5,97	0,6919	
15	58,5 - 59,5	7	7,10	0,0013	
16	59,5 - 60,5	10	9,73	0,0072	
17	60,5 - 61,5	9	10,57	0,2344	
18	61,5 - 62,5	8	9,10	0,1320	
19	62,5 - 63,5	7	6,20	0,1046	
20	63,5 - 65,5	6	4,77	0,3184	
21					
22	Наблюдаемое значение хи-квадрат		Критическое значение хи-квадрат		
23	1,4898 <		9,4877	Нет оснований отвергнуть H0	
24	p-значение		уровень значимости		
25	0,9602 >		0,05	Нет оснований отвергнуть H0	

Рисунок 2.25 – Расчеты для критерия Пирсона по выборке веса учеников

В ячейках I14:I20 рассчитаем элементы суммы по формуле (2.13), введя в ячейку I14 формулу  $= (G14 - H14)^2 / H14$  и скопировав ее вниз по столбцу.

Наблюдаемое значение критерия в ячейке F23 рассчитывается как сумма по столбцу I. Критическое значение критерия в ячейке H23 рассчитаем по формуле  $= \text{ХИ2ОБР}(0,05; 4)$ . При этом принятый уровень значимости ( $\alpha$ ) равен 0,05, а число степеней свободы равно:  $4 = 7 - 2 - 1$  (учтем, что имеем семь интервалов

группировки). Поскольку полученное наблюдаемое значение меньше, чем критическое, делаем вывод, что нельзя отвергнуть нулевую гипотезу о нормальном законе распределения.

Рассчитаем также  $p$ -значение в ячейке F25 с помощью функции

$$=ХИ2ТЕСТ(Г14:Г20;Н14:Н20).$$

Аргументы этой функции следующие:

- *фактический интервал* – наблюдаемые абсолютные частоты (Г14:Г20);
- *ожидаемый интервал* – теоретические абсолютные частоты (Н14:Н20).

Рассчитанное значение в ячейке F25 равно 0,96. Это значительно больше уровня значимости ( $\alpha = 0,05$ ), следовательно, гипотезу о соответствии исследуемого распределения нормальному нужно принять. Кроме того, это значение близко к единице, что говорит о высокой степени этого соответствия.

**Критерий Колмогорова-Смирнова.** Если критерий Пирсона сравнивает эмпирические и теоретические частоты, то суть критерия Колмогорова-Смирнова состоит в сравнении эмпирической и теоретической функции распределения. Для использования этого критерия не требуется какая-либо группировка данных и определение интервалов не вызывает сложности. Кроме того, этот критерий точно достоверен для любого объема выборки (если известны все параметры) в отличие от асимптотически достоверного критерия хи-квадрат.

Недостатком критерия Колмогорова-Смирнова является то, что область его применения более ограничена. Это выражается в том, что исходная форма критерия достоверна только тогда, когда известны параметры распределения и распределение является непрерывным, т. е. параметры не могут оцениваться по данным. Последнее время появились модификации этого критерия, которые позволяют использовать его и в том случае, когда параметры распределения оцениваются по выборке (для случая нормального и экспоненциального распределения)<sup>1</sup>. При этом критические точки такого модифицированного распределения сведены в относительно небольшую таблицу (таблица 2.7).

Статистика, которая лежит в основе критерия Колмогорова-Смирнова – это наибольшее (вертикальное) расстояние между эмпирической функцией распределения  $F_n(x)$  и теоретической функцией распределения  $F(x)$ :

$$D_n = \sup \left\{ \left| F_n(x) - F(x) \right| \right\}.$$

Супремум здесь используется вместо максимума, поскольку максимум может быть не определен. Эмпирическая функция распределения  $F_n(x)$ , как известно, имеет ступенчатый вид, причем на каждом интервале его левый конец не включается (рисунок 2.26). Наибольшее расхождение между теоретической и эмпирической функцией распределения в каждой точке находится как максимум из отклонений до верхнего и нижнего концов «ступеньки»:

$$d_i^1 = |F(x_i) - F_n(x_i)|; \quad d_i^2 = |F_n(x_{i+1}) - F(x_i)|; \quad d_i = \max \{d_i^1, d_i^2\}.$$

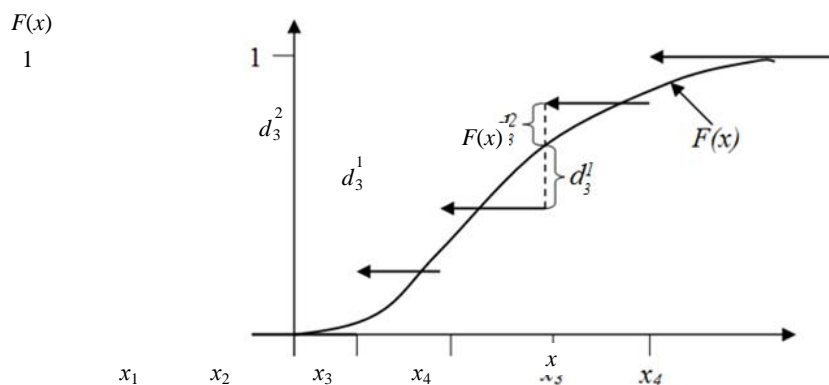


Рисунок 2.26 – Отклонение между теоретической и эмпирической функциями распределения

Таким образом, величина  $D_n$  определяется как наибольшее из отклонений в каждой точке:

$$D_n = \max_i d_i = \max_i \{d_i^1, d_i^2\}.$$

В зависимости от имеющейся информации о параметрах теоретической функции распределения применяется одна из модифицированных статистик (таблица 2.7).

<sup>1</sup> Расчеты для распределения Вейбулла и лог-логистического распределения описаны в книге В. Кельтон, А. Лоу «Имитационное моделирование. Классика CS» (С. 422–423).

Таблица 2.7 – Критические значения для модифицированного критерия Колмогорова-Смирнова

Случай	Статистика	Уровень значимости $\alpha$				
		0,15	0,10	0,05	0,025	0,01
Все параметры известны	$\left(\sqrt{n} + 0,12 + \frac{0,11}{\sqrt{n}}\right) D_n$	1,138	1,224	1,358	1,480	1,628
Нормальный закон распределения $N(\bar{x}, s^2)$	$\left(\sqrt{n} + 0,01 + \frac{0,85}{\sqrt{n}}\right) D_n$	0,775	0,819	0,895	0,955	1,035
Экспоненциальный закон распределения	$\left(D_n - \frac{0,2}{n}\right) \left(\sqrt{n} + 0,26 + \frac{0,5}{\sqrt{n}}\right)$	0,926	0,990	1,094	1,190	1,308

Если известны все параметры теоретической функции распределения (т. е. они не оценивались по данным), то рассчитывается статистика  $\left(\sqrt{n} + 0,12 + \frac{0,11}{\sqrt{n}}\right) D_n$ . Критические точки для различных уровней значимости в этом случае представлены в первой строке таблицы 2.7.

Если предполагается нормальное распределение, параметры которого оценивались по выборке ( $\bar{x}$  и  $s$ ), то рассчитывается статистика  $\left(\sqrt{n} + 0,01 + \frac{0,85}{\sqrt{n}}\right) D_n$ . А если предполагается экспоненциальное распределение, причем в качестве оценки параметра  $\lambda$  использовалось значение  $1/\bar{x}$ , то рассчитывается статистика  $\left(D_n - \frac{0,2}{n}\right) \left(\sqrt{n} + 0,26 + \frac{0,5}{\sqrt{n}}\right)$ . Критические точки показаны во второй и третьей строках таблицы 2.7 соответственно.

Гипотеза  $H_0$  о соответствии эмпирических данных выбранному закону распределения отвергается, если рассчитанное по выборке значение статистического критерия окажется больше критического значения.

**Пример 2.7.** На основании данных о весе студентов из таблицы 2.5 (см. пример 2.3) с помощью критерия Колмогорова-Смирнова проверить гипотезу о нормальном законе распределения исследуемой случайной величины.

#### Решение

Скопируем данные наблюдений на чистый лист Excel и построим вариационный ряд аналогично тому, как это делалось в примере 2.3 (можно использовать функцию ЧАСТОТА() или инструмент Гистограмма из Пакета анализа). Абсолютные частоты вариант расположим в столбце Абсолютные частоты, как показано на рисунке 2.27.

Рассчитаем выборочные среднее и стандартное отклонение, используя функции СРЗНАЧ() и СТАНДОТКЛОН().

В ячейке G11 найдем число всех наблюдений, просуммировав фактические частоты.

В столбцах Н и I рассчитаем относительные и накопленные частоты, как это было описано в примере 2.3. Как отмечалось ранее, накопленные частоты (точки кумулятивной кривой) – это левые концы «ступенек» эмпирической функции распределения, т. е.  $F_n(x_{i+1})$ .

В столбце J рассчитаем теоретические значения функции распределения для каждой варианты ( $F(x_i)$ ). Для этого в ячейку J2 введем формулу =НОРМРАСП(F2;SD\$3;SD\$4;1), которую затем методом автозаполнения скопируем вниз по столбцу. При расчете теоретической функции распределения мы используем оценки математического ожидания и среднеквадратического отклонения, полученные по выборке.

	A	B	C	D	E	F	G	H	I	J	K	L
						Варианты $x_i$	Абсолютные частоты	Относительные частоты	Накопленные частоты $F_n(x_{i+1})$	Теоретическая функция распределения $F(x_i)$	Модуль разности $F(x_i) - F_n(x_{i+1})$	Модуль разности $F(x_i) - F_n(x_i)$
1	наблюдения	максимум	65									
2	64	минимум	57			57		2	0,036	0,036	0,030	0,030
3	57	среднее	60,855			58		6	0,109	0,145	0,082	0,064
4	63	станд. отклон	2,050			59		7	0,127	0,273	0,183	0,090
5	62					60		10	0,182	0,455	0,338	0,116
6	58					61		9	0,164	0,618	0,528	0,090
7	61					62		8	0,145	0,764	0,712	0,052
8	63					63		7	0,127	0,891	0,852	0,039
9	60					64		4	0,073	0,964	0,938	0,026
10	60					65		2	0,036	1,000	0,978	0,022
11	61					Всего наблюдений		55				
12	65											
13	62					Наибольшее отклонение		0,116				
14	62					Наблюдаемое значение статистики		0,876129758				
15	60							<				
16	64					Критическая точка для $\alpha=0,05$		0,895				
17	61											
18	59					Нет оснований отвергнуть гипотезу $H_0$						

Рисунок 2.27 – Расчеты для критерия Колмогорова-Смирнова по выборке веса учеников



ните степень соответствия эмпирических данных нормальному закону распределения с помощью критерия Пирсона. Аналогично проверьте, соответствуют ли эмпирические данные экспоненциальному закону распределения.

6. Возраст студентов одного потока представляется следующими данными (лет): 17, 20, 18, 19, 18, 17, 20, 21, 24, 22, 20, 21, 20, 19, 18, 20, 21, 22, 25, 20.

Постройте вариационный ряд, полигон частот, диаграмму относительных и накопленных частот. Определите выборочные характеристики. Постройте 95%-ный доверительный интервал для математического ожидания. Рассчитайте, сколько необходимо иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,5. Оцените степень соответствия эмпирических данных нормальному закону распределения с помощью критерия Колмогорова-Смирнова.

7. Служба контроля качества предприятия произвела замеры длин (в сантиметрах) случайно отобранных заготовок: 39, 41, 40, 43, 41, 44, 42, 41, 41, 43, 42, 39, 40, 42, 43, 42, 41, 39, 42, 42, 41, 42, 40, 41, 43, 41, 39, 40, 41, 40.

Постройте вариационный ряд, полигон частот, диаграмму относительных и накопленных частот. Определите выборочные характеристики. Постройте 99%-ный доверительный интервал для математического ожидания. Рассчитайте, сколько необходимо иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,3. Оцените степень соответствия эмпирических данных нормальному закону распределения с помощью критериев Пирсона и Колмогорова-Смирнова.

8. Для исследования системы массового обслуживания измерялись интервалы времени (в минутах) между поступлением заявок в систему: 1, 6, 12, 7, 1, 12, 1, 2, 8, 4, 3, 13, 1, 5, 5, 10, 2, 2, 2, 4, 3, 11, 2, 11, 3, 4, 5, 7, 6, 6, 9, 10, 1, 3, 1, 2.

Постройте вариационный ряд, полигон частот, диаграмму относительных и накопленных частот. Определите выборочные характеристики. Постройте 95%-ный доверительный интервал для математического ожидания. Рассчитайте, сколько необходимо иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,5. Оцените степень соответствия эмпирических данных экспоненциальному закону распределения с помощью критериев Пирсона и Колмогорова-Смирнова.

### 3. АНАЛИЗ НЕСКОЛЬКИХ ВЫБОРОК

#### 3.1. Выявление достоверности различий между двумя выборками

Пусть  $X$  и  $Y$  – нормально распределенные случайные величины, и пусть  $\{x_1, x_2, \dots, x_n\}$  является выборкой значений случайной величины  $X$ , а  $\{y_1, y_2, \dots, y_m\}$  – выборкой значений случайной величины  $Y$ . Необходимо по данным выборкам проверить гипотезу о равенстве (однородности) параметров распределений (математических ожиданий  $M(X) = M(Y)$  или дисперсий  $D(X) = D(Y)$ ). Такая проверка выполняется на основании статистических критериев, называемых критериями различия.

**Проверка гипотезы о равенстве дисперсий нормальных генеральных совокупностей.** На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, инструментов, самих методов измерений и т. д. Очевидно, предпочтительнее тот прибор (инструмент и метод), который обеспечивает наименьшее рассеяние результатов измерений, т. е. наименьшую дисперсию.

Пусть генеральные совокупности  $X$  и  $Y$  распределены нормально. По независимым выборкам с объемами, равными соответственно  $n$  и  $m$ , извлеченным из этих совокупностей, найдены выборочные дисперсии  $s_X^2$  и  $s_Y^2$ . Требуется при заданном уровне значимости  $\alpha$  проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии равны между собой:

$$H_0: D(X) = D(Y).$$

Такая задача ставится потому, что обычно выборочные дисперсии оказываются различными. Для определенности пусть  $s_X^2 > s_Y^2$ , в противном случае эти величины можно переобозначить. Возникает вопрос: значимо (существенно) или незначимо (несущественно) различаются выборочные дисперсии?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные дисперсии одинаковы, то различие выборочных дисперсий незначимо и объясняется случайными причинами, в частности, случайным отбором объектов выборки. Например, если различие выборочных дисперсий результатов измерений, выполненных двумя приборами, оказалось незначимым, то приборы имеют одинаковую точность.

Если нулевая гипотеза будет отвергнута, то различие выборочных дисперсий значимо и не может быть объяснено случайными причинами, а является следствием того, что сами генеральные дисперсии различны. Например, если различие выборочных дисперсий результатов измерений, выполненных двумя приборами, оказалось значимым, то точность приборов различна.

В качестве критерия для проверки гипотезы о равенстве генеральных дисперсий принимают отношение большей выборочной дисперсии к меньшей, т. е. случайную величину

$$F = \frac{s_X^2}{s_Y^2}, \text{ если } s_X^2 > s_Y^2. \quad (3.1)$$

Величина  $F$  при условии справедливости нулевой гипотезы имеет распределение Фишера со следующими степенями свободы:  $\nu_1 = n - 1$  и  $\nu_2 = m - 1$ .

Вычисляется наблюдаемое значение статистики ( $F_{набл}$ ). Критическая область строится в зависимости от вида конкурирующей гипотезы:

1. При  $H_1: D(X) > D(Y)$  строят правостороннюю критическую область так, чтобы вероятность попадания критерия  $F$  в эту область при условии справедливости нулевой гипотезы была равна принятому уровню значимости:  $P(F > F_{\alpha}) = \alpha$  (рисунок 3.1).

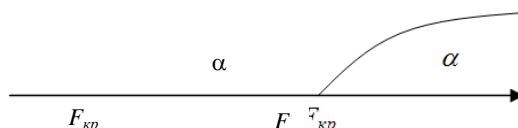


Рисунок 3.1 – Правосторонняя критическая область

В MS Excel для расчета критической точки можно использовать функцию  $FРАСПОБР(\alpha; \nu_1; \nu_2)$ . Если  $F_{набл} \geq F_{\alpha}$ , то гипотеза  $H_0$  отвергается в пользу гипотезы  $H_1$ . Если  $F_{набл} < F_{\alpha}$ , то нет оснований отвергнуть нулевую гипотезу. Кроме того, можно рассчитать  $p$ -значение с помощью функции Excel  $FРАСП(F_{набл}; \nu_1; \nu_2)$ . Если эта величина окажется больше уровня значимости  $\alpha$ , то это свидетельствует в пользу нулевой гипотезы о равенстве генеральных дисперсий.

2. При  $H_1: D(X) \neq D(Y)$  строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область при условии справедливости нулевой гипотезы была равна принятому уровню значимости  $\alpha$ . Можно доказать, что наибольшая мощность критерия достигается тогда, когда вероятность попадания критерия в каждый из двух интервалов критической области равна  $\alpha/2$  (рисунок 3.2), т. е. должны иметь место соотношения:  $P(F > F_{\alpha/2}) = \alpha/2$  и  $P(F < F_{\alpha/2}) = \alpha/2$ .

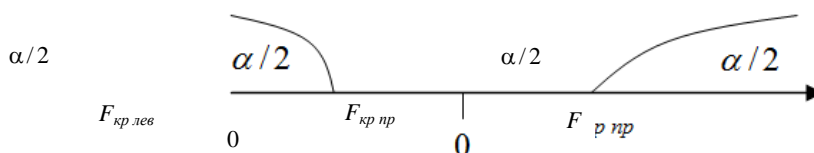


Рисунок 3.2 – Двусторонняя критическая область

Для определения правой критической точки можно использовать функцию MS Excel  $FРАСПОБР(\alpha/2; \nu_1; \nu_2)$ . Левая критическая точка определяется по формуле  $FРАСПОБР(1-\alpha/2; \nu_1; \nu_2)$ . Если  $F_{набл} < F_{\alpha/2}$  или  $F_{набл} > F_{1-\alpha/2}$ , то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевая гипотеза отвергается.  $P$ -значение для двусторонней критической области рассчитывается с помощью функции  $ФТЕСТ(массив1; массив2)$ , где *массив1* – диапазон ячеек, содержащий значения первой выборки, а *массив2* – диапазон, содержащий значения второй выборки. Если рассчитанное  $p$ -значение больше уровня значимости, то нет оснований отвергнуть нулевую гипотезу.

В MS Excel можно также использовать инструмент *Двухвыборочный F-тест для дисперсии* надстройки *Пакет Анализа*. При этом следует иметь ввиду два существенных момента:

- В качестве первой переменной всегда следует выбирать диапазон значений, выборочная дисперсия которого наибольшая.
- Данный инструмент рассчитывает критическую точку для односторонней критической области. Можно показать, что для случая двусторонней критической области достаточно найти  $F_{\alpha/2}$  и проверить условие  $F_{набл} < F_{\alpha/2}$ . Таким образом, и для альтернативной гипотезы  $H_1: D(X) \neq D(Y)$  можно использовать данный инструмент, если задать уровень значимости  $\alpha/2$ .

**Пример 3.1.** Пусть случайная величина  $X$  характеризует расход сырья при производстве продукции по одной технологии,  $Y$  – по другой технологии, причем предполагается, что  $X$  и  $Y$  нормально распределены. В результате наблюдений получены выборки значений случайных величин  $X$  и  $Y$  (таблицы 3.1 и 3.2).

Таблица 3.1 – Выборка значений  $X$

114	112	132	124	119	124	119	116	129	116
124	119	119	114	129	116	124	129	116	119
110	124	140	119	124	129	119	124	124	124
116	129	119	124	110	124	112	114	129	116
119	116	129	116	119	114	132	119	124	112

Таблица 3.2 – Выборка значений  $Y$

109	137	111	133	111	126	111	114	114	114
119	122	119	122	122	122	119	122	122	122
114	119	114	114	119	119	114	119	119	119
122	123	122	123	123	123	122	123	123	123
123	133	123	111	126	126	126	126	126	126
133	114	133	133	111	135	133	135	135	137

Требуется при заданном уровне значимости  $\alpha = 0,05$  проверить гипотезу о равенстве дисперсий генеральных совокупностей  $X$  и  $Y$  ( $H_0: D(X) = D(Y)$ ).

*Решение.*

Введем выборочные данные в столбцы А и В листа MS Excel, озаглавив их соответственно «Выборка  $X$ » и «Выборка  $Y$ » (рисунок 3.3). С помощью функции ДИСП() рассчитаем выборочные дисперсии  $s_X^2$  и  $s_Y^2$  по этим выборкам. Поскольку  $s_Y^2$  больше, то именно диапазон выборочных значений  $Y$  должен выступать в качестве первой переменной при использовании инструмента анализа *Двухвыборочный F-тест для дисперсии*.

	А	В	С	Д	Е	Ф	Г	Н
1	Выборка X	Выборка Y		Выборочная дисперсия X	Выборочная дисперсия Y			
2	114	109		42,52	53,70			
3	112	137		H1: D(Y) > D(X) односторонняя критическая область				
4	132	111		Двухвыборочный F-тест для дисперсии				
5	124	133						
6	119	111			Выборка Y	Выборка X		
7	124	126		Среднее	122,2166667	120,88		
8	119	111		Дисперсия	53,6980226	42,51591837		
9	116	114		Наблюдения	60	50		
10	129	114		df	59	49		Проверка:
11	116	114		F	1,263009825			1,26301
12	124	119		P(F<=f) одностороннее	0,201061568			0,201062
13	119	122		F критическое одностороннее	1,582744994			1,582745
14	110	110						

Рисунок 3.3 – Исходные данные и результат работы инструмента анализа *Двухвыборочный F-тест для дисперсии*

1. Рассмотрим альтернативную гипотезу  $H_1: D(Y) > D(X)$ . В этом случае необходимо использовать одностороннюю критическую область. Вызовем надстройку *Пакет анализа: Сервис/Анализ данных* и выберем инструмент *Двухвыборочный F-тест для дисперсии*. Пример заполнения диалогового окна для этого инструмента показан на рисунке 3.4.



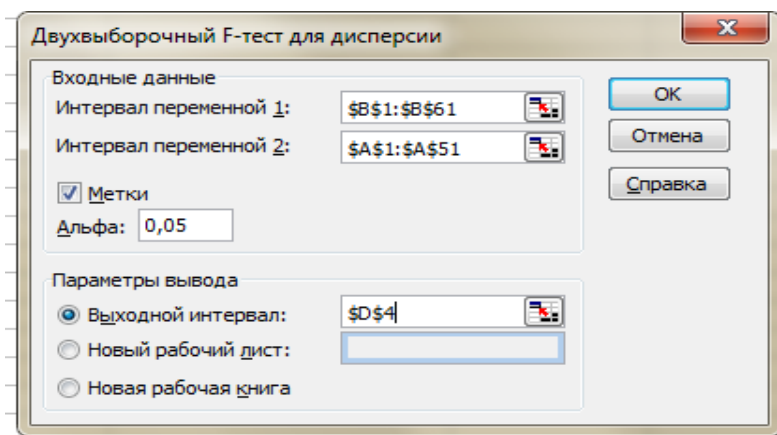


Рисунок 3.4 – Диалоговое окно «Двухвыборочный F-тест для дисперсии» для альтернативной гипотезы  $H_1: D(Y) > D(X)$

Поскольку интервалы переменных указаны вместе с заголовками в первой строке, поставим флажок *Метки*. Это позволит приложению MS Excel использовать данные первой строки в качестве подписей столбцов результирующей таблицы.

Поле *Альфа* содержит значение уровня значимости. Для односторонней критической области берем заданное значение  $\alpha = 0,05$ .

В поле *Выходной интервал* указываем адрес левого верхнего угла результирующей таблицы на этом же листе Excel.

Результат анализа выборок показан на рисунке 3.3. Первые три строки результирующей таблицы содержат выборочные средние, выборочные дисперсии и число наблюдений по каждой выборке соответственно. В строке *df* содержатся значения степеней свободы:  $\nu_1 = n - 1 = 59$  и  $\nu_2 = m - 1 = 49$ . Строка *F* содержит наблюдаемое значение статистики ( $F_{\text{набл}}$ ) отношения большей дисперсии к меньшей. Для проверки рассчитаем это значение самостоятельно, введя в ячейку H11 формулу  $=E8/F8$ .

В строке *F критическое одностороннее* приведено критическое значение ( $F_{\text{кр}}$ ). Для проверки рассчитаем его другим способом в ячейке H13, введя формулу  $=\text{FРАСПОБР}(0,05;E10;F10)$ . Поскольку  $F_{\text{набл}} < F_{\text{кр}}$ , нет оснований отвергнуть нулевую гипотезу о равенстве дисперсий.

В строке *P(F<=f) одностороннее* рассчитано *p*-значение, которое мы также продублируем в ячейке H12, введя формулу  $=\text{FРАСП}(E11;E10;F10)$ . Поскольку найденное *p*-значение больше уровня значимости ( $\alpha = 0,05$ ), то это также свидетельствует в пользу гипотезы о равенстве дисперсий.

2. Рассмотрим альтернативную гипотезу  $H_1: D(X) \neq D(Y)$ . В этом случае критическая область должна быть двусторонней. Как уже отмечалось выше, достаточно найти только правую критическую точку из условия  $P(F > F_{\text{кр прав}}) = \alpha / 2$ . Поэтому снова будем использовать инструмент *Двухвыборочный F-тест для дисперсии*, но в поле *Альфа* зададим следующее значение:  $\alpha / 2 = 0,05 / 2 = 0,025$ . Результат работы этого инструмента статистического анализа показан на рисунке 3.5.

H1: D(Y) ≠ D(X) двусторонняя критическая область		
Двухвыборочный F-тест для дисперсии		
	Выборка Y	Выборка X
Среднее	122,2166667	120,88
Дисперсия	53,6980226	42,51591837
Наблюдения	60	50
df	59	49
F	1,263009825	
P(F<=f) одностороннее	0,201061568	
F критическое одностороннее	1,730442954	
F кр прав	1,730442954	F кр лев
		0,585868694
p-значение двустороннее	0,402123136	

Рисунок 3.5 – Результирующая таблица инструмента *Двухвыборочный F-тест для дисперсии* для альтернативной гипотезы  $D(X) \neq D(Y)$



Ниже таблицы, которую формирует *Пакет анализа*, приведены проверочные расчеты правой и левой критических точек по формулам  $=\text{FРАСПОБР}(0,025;E22;F22)$  и  $=\text{FРАСПОБР}(1-0,025;E22;F22)$  соответственно. Поскольку наблюдаемое значение попадает в область принятия гипотезы ( $0,59 < 1,26 < 1,73$ ), нет оснований отвергнуть гипотезу о равенстве дисперсий и в случае альтернативной гипотезы  $H_1: D(X) \neq D(Y)$ . Отметим, что на практике чаще всего используется именно этот вариант альтернативной гипотезы при сравнении дисперсий двух выборок.  $P$ -значение для двусторонней критической области рассчитано с помощью формулы  $=\text{FТЕСТ}(A2:A51;B2:B61)$ . Очевидно, что эта величина в два раза больше, чем соответствующее значение, рассчитанное *Пакетом анализа* для односторонней критической области. Поскольку  $p$ -значение больше уровня значимости, принимается гипотеза о равенстве дисперсий случайных величин  $X$  и  $Y$ , для которых получены две выборки значений.

**Проверка гипотезы о равенстве математических ожиданий нормальных генеральных совокупностей.** При анализе многих экономических показателей приходится сравнивать две средние величины. Например, можно сравнивать уровни жизни в двух странах по размеру дохода на душу населения, два варианта инвестирования – по размерам средних дивидендов, качество знаний студентов двух университетов – по среднему баллу на комплексном тестовом экзамене. Однако следует учитывать, что различие средних величин, рассчитанных по выборочным данным, не означает автоматически различие и математических ожиданий соответствующих генеральных совокупностей. Для корректного сравнения требуется проверка гипотезы о равенстве математических ожиданий.

Пусть генеральные совокупности  $X$  и  $Y$  распределены нормально. По независимым выборкам с объемами, равными соответственно  $n$  и  $m$ , извлеченным из этих совокупностей, найдены выборочные средние  $\bar{x}$  и  $\bar{y}$ , а также выборочные дисперсии  $s_x^2$  и  $s_y^2$ . Требуется проверить при заданном уровне значимости  $\alpha$  нулевую гипотезу, состоящую в том, что математические ожидания равны между собой:  $H_0: M(X) = M(Y)$ .

Как правило, выборочные средние оказываются различными. Возникает вопрос: значимо (существенно) или незначимо (несущественно) различаются выборочные средние?

Если окажется, что нулевая гипотеза справедлива, т. е. математические ожидания генеральных совокупностей равны, то различие выборочных средних незначимо и объясняется случайным отбором объектов выборки. Например, сравниваются результаты централизованного тестирования выпускников средних школ в двух городах. Если оказывается, что средние баллы по математике различаются незначимо, то это говорит об одинаковом уровне образования в средних учебных заведениях этих городов.

Если же нулевая гипотеза отвергнута, т. е. математические ожидания различны, то различие выборочных средних значимо и не может быть объяснено случайными причинами. Например, в случае о результатах тестирования это бы означало, что в одном из городов лучше обучают математике в силу каких-либо причин (наличие специализированных классов, более квалифицированных преподавателей и т. п.).

Выбор статистического критерия и, следовательно, инструмента проверки гипотезы о равенстве математических ожиданий в MS Excel зависит от следующих факторов:

- являются ли выборки независимыми или они зависимы;
- является ли закон распределения случайных величин  $X$  и  $Y$  нормальным или он существенно отличается от нормального;
- каков (большой или малый) объем каждой выборки;
- известны ли дисперсии случайных величин  $X$  и  $Y$ ;
- если дисперсии неизвестны, то можно ли считать их равными, или же они не равны.

В таблице 3.3 приведены варианты сочетания этих факторов и указано, какой инструмент *Пакета анализа* и какие стандартные функции Excel следует использовать в каждом случае.

Выборки называются *независимыми*, если данные наблюдений не связаны между собой. Например, проверяется эффективность какого-либо лекарства, и есть опытная группа людей (которые принимают лекарство) и контрольная группа, состоящие из разных пациентов. При этом количество наблюдений в выборках может быть различно.

В случае *зависимых* выборок одна и та же группа объектов порождает числовой материал для выборок. Например, измеряется содержание лейкоцитов у здоровых животных, а затем у тех же животных, подвергнутых некоторой дозе радиационного облучения. В этом случае размеры выборок должны быть одинаковыми. Для проверки гипотезы о равенстве математических ожиданий зависимых выборок используется  $t$ -критерий Стьюдента (первая строка таблицы 3.3). Он реализован в инструменте *Парный двухвыборочный  $t$ -тест для средних* из *Пакета анализа*.

Как правило, для использования статистических критериев, показанных в таблице 3.3, требуется, чтобы закон распределения случайных величин  $X$  и  $Y$ , из которых получены выборки, был нормальным (или хотя бы близок к нему). Для проверки этого условия используются критерии Пирсона и Колмогорова-Смирнова (см. подраздел 2.4). Если же закон распределения для сравниваемых выборок существенно отличается от нормального, то проверить гипотезу о равенстве математических ожиданий можно только в том случае, если объем выборок достаточно велик (вторая строка таблицы 3.3). Следует отметить, что в научной литературе существуют разные мнения о том, какой объем выборки считать большим. Одни авторы называют большой выборку при  $n > 30$ , другие – при  $n > 60$  и т. д.

Если дисперсии генеральных совокупностей  $X$  и  $Y$  известны (например, из предшествующего опыта или найдены теоретически), то используется статистический критерий  $Z$ , имеющий стандартный нормальный закон распределения (строка 3 таблицы 3.3).

Этот же критерий можно использовать и в случае выборок большого объема, если заменить точные значения дисперсий на их выборочные аналоги. Полученный таким образом критерий является приближенным (вторая строка таблицы 3.3).

Таблица 3.3 – Выбор инструмента для проверки гипотезы о равенстве математических ожиданий двух выборок

Выборки	Распределение	Размер выборок	Дисперсии	Статистика	Инструмент <i>Пакета анализа</i>	Функция MS Excel для расчета критической точки	Функция MS Excel для расчета $p$ -значения
Зависимые	Нормальное	$n = m$	Не известны	Случайная величина $T = \frac{\bar{d}\sqrt{n}}{s_d}, d_i = x_i - y_i$ имеет $t$ -распределение Стьюдента с $k = n - 1$ степенями свободы	Парный двухвыборочный $t$ -тест для средних	Двусторонняя критическая область СТЬЮДРАСПОБР( $\alpha$ ; $k$ ). Односторонняя критическая область СТЬЮДРАСПОБР ( $2\alpha$ ; $k$ )	Двусторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 2). Односторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 1)
Независимые	Произвольное	Большие ( $>30$ )	Не известны	Случайная величина $Z = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}}$ стремится к стандартному нормальному распределению (при больших $n$ и $m$ )	Двухвыборочный $Z$ -тест для средних	Односторонняя критическая область НОРМСТОБР( $1 - \alpha$ ). Двусторонняя критическая область НОРМСТОБР( $1 - \alpha/2$ )	Односторонняя критическая область 1-НОРМРАСП ( $Z_{набл}$ ; 0;1;1). Двусторонняя критическая область 2*[1-НОРМРАСП ( $Z_{набл}$ ; 0;1;1)]
	Нормальное	Малые	Известны	Случайная величина $Z = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/m}}$ имеет стандартное нормальное распределение	Двухвыборочный $Z$ -тест для средних	Односторонняя критическая область НОРМСТОБР( $1 - \alpha$ ). Двусторонняя критическая область НОРМСТОБР( $1 - \alpha/2$ )	Односторонняя критическая область 1-НОРМРАСП ( $Z_{набл}$ ; 0;1;1). Двусторонняя критическая область 2*[1-НОРМРАСП ( $Z_{набл}$ ; 0;1;1)]
		Малые	Не известны, равны	Случайная величина $T = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2} / \sqrt{n+m}} \sqrt{\frac{nm(n+m-2)}{n+m}}$ имеет $t$ -распределение Стьюдента с $k = n + m - 2$ степенями свободы	Двухвыборочный $t$ -тест с одинаковыми дисперсиями	Двусторонняя критическая область СТЬЮДРАСПОБР ( $\alpha$ ; $k$ ). Односторонняя критическая область СТЬЮДРАСПОБР ( $2\alpha$ ; $k$ )	Двусторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 2). Односторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 1)
		Малые	Не известны, не равны	Случайная величина $T = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}}$ имеет $t$ -распределение Стьюдента	Двухвыборочный $t$ -тест с различными дисперсиями	Двусторонняя критическая область СТЬЮДРАСПОБР ( $\alpha$ ; $k$ ). Односторонняя критическая область СТЬЮДРАСПОБР ( $2\alpha$ ; $k$ )	Двусторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 2). Односторонняя критическая область СТЬЮДРАСП( $t_{набл}$ ; $k$ ; 1)

При малых объемах выборок и неизвестных дисперсиях нужно установить, можно ли считать их одинаковыми (например, проверить гипотезу о равенстве дисперсий с помощью  $F$ -критерия, как показано выше). При равных и неравных дисперсиях используются различные формы критерия Стьюдента (четвертая и пятая строки таблицы 3.3), а в *Пакете анализа* это реализовано с помощью инструментов *Двухвыборочный  $t$ -тест с одинаковыми дисперсиями* и *Двухвыборочный  $t$ -тест с различными дисперсиями*.

Кроме процедур *Пакета анализа*, в приложении MS Excel для реализации критерия Стьюдента используется специальная функция ТТЕСТ(), которая имеет следующий формат: ТТЕСТ(*массив\_1*; *массив\_2*; *хвосты*; *тип*), где:

- *массив\_1* – элементы первой выборки;
- *массив\_2* – элементы второй выборки;
- *хвосты* – число хвостов распределения (два, если строится двусторонняя критическая область, один – если односторонняя);
- *тип* – вид исполняемого теста (если *тип* = 1, то реализуется парный  $t$ -критерий (зависимые выборки), если *тип* = 2 – двухвыборочный тест с равными дисперсиями, если *тип* = 3 – двухвыборочный тест с неравными дисперсиями).

Рассмотрим процедуру проверки гипотезы о равенстве математических ожиданий для случая, когда случайные величины  $X$  и  $Y$  распределены нормально, а их дисперсии известны. Остальные случаи рассматриваются аналогично.

Итак, пусть требуется проверить нулевую гипотезу  $H_0: M(X) = M(Y)$ . По независимым выборкам из генеральных совокупностей  $X$  и  $Y$ , объемы которых равны соответственно  $n$  и  $m$ , найдены выборочные средние  $\bar{x}$  и  $\bar{y}$ .

В качестве критерия проверки нулевой гипотезы принимается случайная величина  $Z = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/m}}$ , имеющая стандартный нормальный закон распределения. По выборочным данным рассчитывается наблюдаемое значение критерия ( $Z_{набл}$ ).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Рассмотрим два случая построения критической области:

1. Конкурирующая гипотеза  $H_1: M(X) \neq M(Y)$ .

В этом случае строится двусторонняя критическая область так, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости  $\alpha$ . Наибольшая мощность критерия достигается тогда, когда «левая» и «правая» критические точки выбраны так, что вероятность попадания критерия в каждый из двух интервалов критической области равна  $\alpha/2$  (рисунок 3.6).

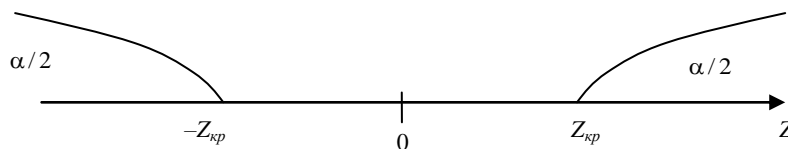


Рисунок 3.6 – Двусторонняя критическая область для Z-критерия

Поскольку  $Z$  – стандартная нормальная величина, а распределение такой величины симметрично относительно нуля, то критические точки симметричны относительно нуля, т. е. если обозначить правую критическую точку  $Z_{кр}$ , то левая критическая точка равна  $-Z_{кр}$ .

В MS Excel для расчета критической точки можно использовать функцию НОРМСТОБР( $1-\alpha/2$ ).

Если  $|Z_{\text{факт}}| < Z_{\text{кр}}$ , то нет оснований отвергнуть нулевую гипотезу.

Если  $|Z_{\text{факт}}| > Z_{\text{кр}}$ , то нулевую гипотезу отвергают.

2. Конкурирующая гипотеза  $H_1: M(X) > M(Y)$ .

На практике такой случай имеет место, если профессиональные соображения позволяют предположить, что математическое ожидание одной случайной величины больше математического ожидания другой. Например, если введено усовершенствование технологического процесса, то естественно допустить, что оно приведет к увеличению выпуска продукции. В этом случае строят правостороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости  $P(Z > Z_{\text{кр}}) = \alpha$  (рисунок 3.7).



Рисунок 3.7 – Правосторонняя критическая область для Z-критерия

В MS Excel для расчета критической точки в этом случае используется функция НОРМСТОБР( $1-\alpha$ ).

Если  $Z_{\text{факт}} < Z_{\text{кр}}$ , то нет оснований отвергнуть нулевую гипотезу.

Если  $Z_{\text{факт}} > Z_{\text{кр}}$ , то нулевую гипотезу отвергают.

Отметим, что конкурирующая гипотеза  $H_1: M(X) < M(Y)$  в данном пособии рассматриваться не будет, поскольку ее всегда можно заменить гипотезой  $H_1: M(X) > M(Y)$ , если переобозначить случайные величины. Кроме того, инструменты *Пакета анализа* ориентированы именно на вид конкурирующей гипотезы  $H_1: M(X) > M(Y)$ . Поэтому при использовании этих инструментов в качестве первой переменной всегда нужно задавать ту случайную величину, у которой выборочное среднее больше (т. е. есть основание предположить, что и математическое ожидание тоже будет больше).

Кроме того, инструменты *Пакета анализа* позволяют проверить не только гипотезу о равенстве математических ожиданий, но и гипотезу о том, что они отличаются на некоторую известную величину. Эта величина задается в диалоговом окне как *Гипотетическая разность средних*. Таким образом,

MS Excel использует критерий  $Z = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{D(X)/n + D(Y)/m}}$ , где  $\delta$  – гипотетическая разность средних.

**Пример 3.2.** Для случайных величин  $X$  и  $Y$ , выборки значений которых показаны в таблицах 3.1 и 3.2, проверить гипотезу о равенстве математических ожиданий при уровне значимости 0,05.

*Решение*

Введем выборочные данные в столбцы А и В листа MS Excel, озаглавив их соответственно «Выборка Х» и «Выборка Y» (рисунок 3.8). С помощью функций СРЗНАЧ() и ДИСП() рассчитаем выборочные средние и дисперсии.

	А	В	С	Д	Е
1	Выборка X	Выборка Y		X	Y
2	114	109	Выборочное среднее	120,88	122,22
3	112	137	Выборочная дисперсия	42,52	53,70
4	132	111			
5	124	133	Двухвыборочный z-тест для средних		
6	119	111			
7	124	126		Выборка Y	Выборка X
8	119	111	Среднее	122,2166667	120,88
9	116	114	Известная дисперсия	53,7	42,52
10	129	114	Наблюдения	60	50
11	116	114	Гипотетическая разность средних	0	
12	124	119	z	1,011755635	
13	119	122	P(Z<=z) одностороннее	0,155827453	
14	119	119	z критическое одностороннее	1,644853627	
15	114	122	P(Z<=z) двухстороннее	0,311654907	
16	129	122	z критическое двухстороннее	1,959963985	

Рисунок 3.8 – Исходные данные и результаты двухвыборочного Z-теста

Очевидно, что данные выборки не могут быть зависимыми. Во-первых, это данные по *разным* технологиям, а не по одной технологии, например, до и после усовершенствования. Во-вторых, выборки имеют разный объем. Объемы выборок довольно велики ( $n = 50$  и  $m = 60$ ), поэтому можно не утруждать себя проблемой доказательства того, что соответствующие случайные величины имеют нормальный закон распределения (и не анализировать равенство дисперсий), а использовать приближенный Z-критерий (вместо дисперсий использовать их выборочные аналоги). При этом получено:  $\bar{x} = 120,88 < \bar{y} = 122,22$ . Поэтому в диалоговом окне *Пакета анализа* первой переменной будет диапазон в столбце В (мы предполагаем в качестве альтернативной гипотезы, что соответствующее математическое ожидание больше).

Зададим команду *Сервис/Анализ данных* и выберем в качестве инструмента *Двухвыборочный Z-тест для средних*. Заполним диалоговое окно, как показано на рисунке 3.9.

*Гипотетическая разность средних* установлена равной нулю, так как проверяется гипотеза о равенстве математических ожиданий, а не их отличии на заданную величину. Дисперсии нужно ввести с клавиатуры (нельзя сделать ссылку на ячейки), поэтому используется их округленное значение. Флажок *Метки* установлен потому, что диапазоны переменных включают заголовки в первой строке.

Рисунок 3.9 – Диалоговое окно «Двухвыборочный Z-тест для средних»

Результат показан на рисунке 3.8. По каждой выборке рассчитано среднее значение и число наблюдений (дисперсии не рассчитывались, а были заданы). В строке  $z$  показано наблюдаемое значение Z-критерия ( $Z_{набл}$ ). Показаны также значения критических точек для альтернативной гипотезы  $M(X) \neq M(Y)$  ( $z$  критическое двустороннее) и для альтернативной гипотезы  $M(X) > M(Y)$  ( $z$  критическое одностороннее). Причем в данном случае присутствует переобозначение переменных (на самом деле конкурирующая гипотеза  $M(Y) > M(X)$ ). Поскольку  $|1,01| < 1,96$  и  $1,01 < 1,64$ , то нет оснований отвергнуть гипотезу о равенстве математических ожиданий ни для двусторонней, ни для односторонней критической области.

Кроме того, были найдены  $p$ -значения для двух видов конкурирующих гипотез.  $P(Z \leq z)$  одностороннее приблизительно равно 0,16. Это означает, что нулевая гипотеза должна была бы быть принята для всех уровней значимости, меньших 0,16 (при односторонней критической области).  $P(Z \leq z)$  двустороннее равно 0,31. Это свидетельствует о том, что для двусторонней критической области нулевая гипотеза принимается при всех уровнях значимости, меньших 0,31. Таким образом, имеем еще одно основание принять нулевую гипотезу при уровне значимости 0,05.

Для проверки и лучшего усвоения материала можно получить все результаты и с помощью функций MS Excel, введя формулы как показано на рисунке 3.10.

	F
12	= (D8-E8)/КОРЕНЬ(D9/D10+E9/E10)
13	=1-НОРМРАСП(D12;0;1;1)
14	=НОРМСТОБР(1-0,05)
15	=2*(1-НОРМРАСП(D12;0;1;1))
16	=НОРМСТОБР(1-0,025)

Рисунок 3.10 – Расчеты для Z-критерия с помощью функций MS Excel

**Пример 3.3.** Сравнивается количество наличных денег у группы студентов до и после посещения буфета (таблица 3.4). Определите достоверность различий между этими группами путем сравнения математических ожиданий генеральных совокупностей при уровне значимости 0,05.

Таблица 3.4 – Данные о количестве денег у студентов, тыс. р.

До буфета	3	3	4	5	6
После буфета	1	2	3	4	5

#### Решение

Исходные данные занесем на лист MS Excel (рисунок 3.11). Рассчитаем выборочные средние в каждой группе:  $\bar{x} = 4,2$  и  $\bar{y} = 3$ . Поскольку известно, что это одна и та же группа студентов, выборки являются зависимыми. Зададим команду *Сервис/Анализ данных* и выберем инструмент *Парный двухвыборочный t-тест для средних*. Заполним диалоговое окно, как это показано на рисунке 3.12.

	A	B	C	D	E
1	До буфета	После буфета		до буфета	после буфета
2	3	1	среднее	4,2	3
3	3	2	Парный двухвыборочный t-тест для средних		
4	4	3			
5	5	4		До буфета	После буфета
6	6	5	Среднее	4,2	3
7			Дисперсия	1,7	2,5
8			Наблюдения	5	5
9			Корреляция Пирсона	0,9701425	
10			Гипотетическая разность средних	0	
11			df	4	
12			t-статистика	6	
13			P(T<=t) одностороннее	0,00194127	
14			t критическое одностороннее	2,13184678	
15			P(T<=t) двустороннее	0,00388254	
16			t критическое двустороннее	2,77644511	

Рисунок 3.11 – Результаты анализа зависимых выборок

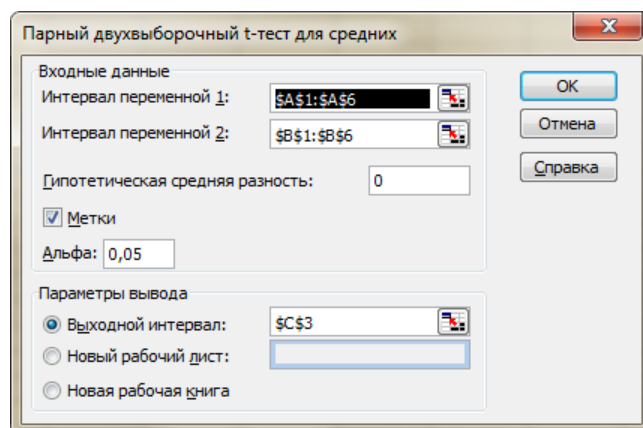


Рисунок 3.12 – Диалоговое окно  
«Парный двухвыборочный  $t$ -тест для средних»

Наблюдаемое значение критерия ( $t_{набл}$ ) равно 6. Поскольку мы хотим убедиться, что после посещения буфета количество денег в кармане студента уменьшается, рассмотрим конкурирующую гипотезу  $H_1: M(X) > M(Y)$ , для которой строится односторонняя критическая область. Критическая точка ( $t_{кр}$ ) равна 2,13 ( $t$  критическое одностороннее). Таким образом,  $t_{набл} > t_{кр}$ , и нулевую гипотезу следует отвергнуть в пользу альтернативной гипотезы  $M(X) > M(Y)$ . Отличия в количестве денег в карманах студентов есть результат посещения буфета. Об этом свидетельствует и  $p$ -значение ( $P(T \leq t)$  одностороннее, равно 0,002). Гипотеза о равенстве математических ожиданий отклоняется, так как  $p < \alpha = 0,05$ .

Другой способ проверить гипотезу о равенстве математических ожиданий – это использовать функцию ТЕСТ. При этом нужно взять  $тип=1$ , так как это зависимые выборки и нужно использовать парный критерий Стьюдента, и  $хвосты=1$ , так как рассматривается односторонняя критическая область. Введем в какую-либо ячейку формулу =ТЕСТ(A2:A6;B2:B6;1;1) и получим такое же  $p$ -значение, как было получено с помощью Пакета анализа ( $p = 0,002$ ).

### Задания для самостоятельной работы

1. Данные о реализации путевок различными филиалами туристической фирмы до и после проведения рекламной кампании представлены в таблице 3.5. Выявите, достоверны ли различия в продаже путевок, проверив гипотезу о равенстве математических ожиданий (уровень значимости 0,05).

Таблица 3.5 – Данные о продаже путевок филиалами туристической фирмы, шт.

Месяц	1-й	2-й	3-й	4-й	5-й	6-й	7-й
Количество проданных путевок:							
до рекламы	135	126	115	140	121	112	130
после рекламы	162	156	144	137	125	145	151

2. Даны результаты бега на 100 м студентов двух групп (таблица 3.6). Студенты первой группы в течение года посещали факультативные занятия по физкультуре. Определите, достоверны ли отличия по результатам бега в этих группах (уровень значимости 0,05).

Таблица 3.6 – Результаты сдачи норматива по бегу, с

Студенты, посещавшие факультатив	12,6	12,3	11,9	12,2	13,0	12,4
Студенты, не посещавшие факультатив	12,8	13,2	13,0	12,9	13,5	13,1

3. Данные о ежемесячной результативности (количество голов) футбольной команды в двух сезонах приведены в таблице 3.7. Определите, есть ли статистические различия в ежемесячной результативности команды в рассматриваемых сезонах (уровень значимости 0,1).

Таблица 3.7 – Количество голов, забитых футбольной командой

Месяц	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь
Количество забитых голов:									
в 2000 г.	3	4	5	8	9	1	2	4	5
в 2001 г.	6	19	3	2	14	4	5	17	1

4. Рассматривается заработная плата обслуживающего персонала гостиницы и работников ресторана (таблица 3.8). Определите, можно ли по этим данным сделать вывод о большей заработной плате работников ресторана (уровень значимости 0,05).

Таблица 3.8 – Данные о заработной плате работников, тыс. р.

Месяц	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й
Размер заработной платы: персонала гостиницы	2 100	2 100	2 000	2 000	2 000	1 900	1 800	1 800
работников ресторана	3 200	2 500	3 000	2 000	1 900	1 800	–	–

5. В таблице 3.9 приведены результаты группы студентов по скоростному чтению до и после посещения специального курса по быстрому чтению. Определите, произошли ли статистические изменения скорости чтения у студентов (уровень значимости 0,1).

Таблица 3.9 – Скорость чтения (знаков в минуту)

Номер студента по списку	1	2	3	4	5	6	7	8	9	10
Скорость чтения: до курса	86	83	86	70	66	90	70	85	77	86
после курса	82	79	91	77	68	86	81	90	85	94

6. Опыты по сравнению веса одного и того же объема азота, полученного после химической очистки из азотистых соединений (X) и из воздуха (Y), проводились при неизменных условиях (15°C и 760 мм. рт. ст.). Результаты измерений представлены в таблице 3.10. Проверьте гипотезу о равенстве дисперсий этих двух случайных величин при уровне значимости 0,05. Укажите, можно ли считать, что оба метода химической очистки дают одинаково точные результаты.

Таблица 3.10 – Вес полученного азота при различных способах химической очистки, г

Номер наблюдений	1	2	3	4	5
Способы химической очистки:					
X	2 301,43	2 298,90	2 298,16	2 301,82	2 298,69
Y	2 310,17	2 309,86	2 310,10	2 310,01	2 310,24

Продолжение таблицы 3.10

Номер наблюдений	6	7	8	9	10
Способы химической очистки:					
X	2 299,40	2 298,40	2 298,89	–	–
Y	2 310,10	2 310,28	2 310,35	2 310,26	2 310,24

7. Производительность двух моторных заводов, выпускающих дизельные двигатели, характеризуется данными, представленными в таблице 3.11. Определите, можно ли считать одинаковой производительность обоих заводов при уровне значимости 0,05.

Таблица 3.11 – Данные о количестве произведенных моторов по месяцам, ед.

Месяц	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	10-й
Количество произведенных моторов:										
1-м заводом	72	84	69	74	82	67	75	86	68	61
2-м заводом	55	65	73	66	58	71	77	68	68	59

8. Сравните точность обработки изделий на двух станках. С этой целью на первом станке было обработано 10, а на втором – 15 изделий. Отклонения контролируемого размера от заданного (в десятых долях миллиметров) показаны в таблице 3.12 (уровень значимости 0,01).

Таблица 3.12 – Отклонение размера детали от заданного (в десятых долях мм)

Номер наблюдений	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Отклонение размера деталей:															
на 1-м станке	2	4	6	2	8	9	4	4	6	5	–	–	–	–	–
на 2-м станке	1	3	2	2	5	7	8	5	5	3	7	8	8	5	9

### 3.2. Дисперсионный анализ

Задачей дисперсионного анализа является изучение влияния одного или нескольких *качественных* факторов на рассматриваемый признак (наблюдаемую случайную величину).

Рассмотрим только однофакторный дисперсионный анализ. Пусть фактор имеет несколько градаций (уровней, групп). Например, требуется выяснить, какой вид удобрений наиболее эффективен для получения хорошего урожая. Тогда фактор – удобрение, его уровни – виды удобрений, а изучаемая случайная величина – урожайность с гектара. Для каждого уровня (вида удобрения) зафиксирована выборка значений урожайности. Причем в общем случае размеры этих выборок могут быть различными.

Таким образом, имеем несколько случайных величин  $X_1, X_2, \dots, X_k$ , где  $k$  – число уровней фактора. Каждая случайная величина  $X_j$  соответствует определенному уровню фактора  $A_j$ , и для нее получена выборка значений  $\{x_{1j}, x_{2j}, \dots, x_{n_jj}\}$ , где  $n_j$  – число наблюдений для данного уровня. Данные наблюдений можно представить в виде таблицы 3.13, в которой количество элементов в каждом столбце может быть различным. При этом  $n = n_1 + n_2 + \dots + n_k$  – общее число всех наблюдений.

Таблица 3.13 – Данные наблюдений для однофакторного дисперсионного анализа

Номер наблюдений	Уровни (группы) фактора			
	$A_1$	$A_2$	...	$A_k$
1	$x_{11}$	$x_{12}$	...	$x_{1k}$
2	$x_{21}$	$x_{22}$	...	$x_{2k}$
...	...	...	...	...
$n_j$	$x_{n_j1}$	$x_{n_j2}$	...	$x_{n_jk}$
Групповая средняя	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$

Требуется проверить гипотезу о равенстве математических ожиданий, соответствующих уровням  $H_0: M(X_1) = M(X_2) = \dots = M(X_k)$ . Другими словами, требуется установить, значимо или незначимо различаются групповые средние.

Попарное сравнение математических ожиданий (см. подраздел 3.1) будет слишком трудоемким. Поэтому для анализа влияния фактора пользуются другим методом, который основан на сравнении дисперсий и поэтому назван *дисперсионным анализом*.

Суть дисперсионного анализа состоит в сравнении дисперсии, которая обусловлена случайными причинами, с дисперсией, вызванной влиянием исследуемого фактора. Если они значимо различаются, то считают, что фактор оказывает влияние на исследуемую величину. Тогда и математические ожидания для уровней будут различаться.

Иногда дисперсионный анализ применяют, чтобы установить однородность нескольких совокупностей. Однородные совокупности можно объединить в одну и тем самым получить о ней более полную информацию и более надежные выводы.

Дисперсионный анализ может быть применен в следующих случаях:

- если генеральные совокупности  $X_1, X_2, \dots, X_k$  распределены нормально и имеют одинаковую, хотя и неизвестную, дисперсию;

- если наблюдения независимы и проводятся в одинаковых условиях.

По данным наблюдений рассчитываются следующие величины:

- *групповые средние*:

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad (j = \overline{1, k}); \quad (3.2)$$

- *общая средняя*:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}.$$

*Общая сумма квадратов отклонений* наблюдаемых значений от общей средней

$$Q_{i\hat{a}\hat{u}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \quad (3.4)$$

может быть представлена в виде суммы:  $Q_{i\hat{a}\hat{u}} = Q_{\hat{o}\hat{a}\hat{e}\hat{o}} + Q_{i\hat{n}\hat{o}}$ .

*Факторная сумма квадратов отклонений* групповых средних от общей средней, которая характеризует рассеяние «между группами» и отражает влияние фактора, определяется по формуле

$$Q_{\hat{o}\hat{a}\hat{e}\hat{o}} = \sum_{j=1}^k (\bar{x}_j - \bar{x})^2 n_j.$$



Остаточная сумма квадратов отклонений наблюдаемых значений группы от своих групповых средних, которая характеризует рассеяние «внутри группы» и отражает влияние случайных причин, выражается формулой

$$Q_{\bar{m}\bar{o}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2.$$

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s_{\bar{i}\bar{a}\bar{u}}^2 = \frac{Q_{\bar{i}\bar{a}\bar{u}}}{n-1}; s_{\bar{o}\bar{a}\bar{e}\bar{o}}^2 = \frac{Q_{\bar{o}\bar{a}\bar{e}\bar{o}}}{k-1}; s_{\bar{m}\bar{o}}^2 = \frac{Q_{\bar{m}\bar{o}}}{n-k}.$$

Если факторная и остаточная дисперсии различаются незначимо, то влияние фактора можно считать незначительным и, следовательно, принять гипотезу о равенстве математических ожиданий уровней  $H_0: M(X_1) = M(X_2) = \dots = M(X_k)$ .

Для проверки различия факторной и остаточной дисперсий используется критерий Фишера:

$$F = \frac{s_{\bar{o}\bar{a}\bar{e}\bar{o}}^2}{s_{\bar{m}\bar{o}}^2} = \frac{\frac{1}{k-1} \sum_{j=1}^k (\bar{x}_j - \bar{x})^2 n_j}{\frac{1}{n-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}.$$

Число степеней свободы критерия Фишера равно:  $\nu_1 = k-1$  и  $\nu_2 = n-k$ .

По выборкам находится наблюдаемое значение критерия ( $F_{\bar{i}\bar{a}\bar{a}\bar{e}}$ ). Строится правосторонняя критическая область, для чего определяется  $F_{\bar{e}\bar{o}}(\alpha, \nu_1, \nu_2)$ . В MS Excel для расчета критической точки можно использовать функцию ФРАСПОБР(). Если  $F_{\bar{i}\bar{a}\bar{a}\bar{e}} < F_{\bar{e}\bar{o}}$ , то можно считать, что факторная и остаточная дисперсии различаются незначимо. Поэтому влияние фактора признается несущественным и нулевая гипотеза о равенстве групповых математических ожиданий не может быть отвергнута.

Проверку гипотезы можно реализовать и с помощью  $p$ -значения. В MS Excel этот показатель рассчитывается с помощью функции ФРАСП ( $F_{\bar{i}\bar{a}\bar{a}\bar{e}}; \nu_1; \nu_2$ ). Если полученное число  $p > \alpha$ , то нулевую гипотезу о равенстве математических ожиданий групп нужно принять (нет влияния фактора). Если же  $p < \alpha$ , то нулевая гипотеза отклоняется, т. е. признается влияние фактора.

Однофакторный дисперсионный анализ в MS Excel также можно реализовать с помощью инструмента *Однофакторный дисперсионный анализ*, который можно выбрать в диалоговом окне по команде *Сервис/Анализ данных*.

Степень влияния фактора на результативный показатель может быть измерена с помощью *выборочного коэффициента детерминации* ( $R^2$ ), показывающего, какая доля общей вариации объясняется влиянием исследуемого фактора:

$$R^2 = \frac{Q_{\bar{o}\bar{a}\bar{e}\bar{o}}}{Q_{\bar{i}\bar{a}\bar{u}}}.$$

**Пример 3.4.** Результаты наблюдений за расходом сырья при производстве одинаковой продукции по одной и той же технологии на пяти различных заводах равных мощностей представлены в таблице 3.14. Известно, что расход сырья является нормально распределенной случайной величиной и дисперсии наблюдений по каждому заводу равны.

При уровне значимости 0,05 требуется выяснить, зависит ли расход сырья от того, на каком заводе произведена продукция.

Таблица 3.14 – Данные наблюдений за расходом сырья на пяти заводах, усл. ед.

Месяцы	Расход сырья				
	1-м заводом	2-м заводом	3-м заводом	4-м заводом	5-м заводом
1-й	114	112	132	124	124
2-й	124	119	124	114	116
3-й	110	124	129	119	119
4-й	116	116	129	124	119
5-й	119	116	129	116	132
6-й	119	124	124	116	129
7-й	129	112	114	129	116
8-й	124	119	119	124	119
9-й	110	119	124	114	–
10-й	124	112	–	116	–
11-й	119	–	–	129	–
12-й	124	–	–	–	–

### Решение

Введем исходные данные на лист MS Excel, как показано на рисунке 3.13. Для каждого завода рассчитаем групповую среднюю. Так, в ячейку A15 введем формулу =СРЗНАЧ(A2:A13), которую затем скопируем методом автозаполнения вправо по строке. При этом Excel игнорирует пустые ячейки при расчете среднего значения (как и при использовании других статистических функций). Общую среднюю рассчитаем с помощью функции СРЗНАЧ(A2:E13).

Аналогично рассчитаем суммы квадратов отклонений от среднего для каждого завода. Введем формулу =КВАДРОТКЛ(A2:A13) в ячейку A19, которую затем скопируем вправо по строке. Общую сумму квадратов отклонений рассчитаем согласно формуле (3.4) с помощью функции КВАДРОТКЛ(A2:E13).

Общее число наблюдений рассчитаем с помощью функции СЧЕТ(A2:E13), которая подсчитывает число заполненных числами ячеек в заданном диапазоне, а пустые ячейки игнорирует.

	A	B	C	D	E	F	G	H	I	J	K	L
	Завод 1	Завод 2	Завод 3	Завод 4	Завод 5		Сумма квадратов отклонений	Число степеней свободы	Дисперсия	Fнабл	p-значение	Fкр
1												
2	114	112	132	124	124	межгрупповая	304,44	4	76,11	2,42	0,063	2,58
3	124	119	124	114	116	внутригрупповая	1417,88	45	31,51			
4	110	124	129	119	119	общая	1722,32					
5	116	116	129	124	119	выборочный коэффициент детерминации	0,18					
6	119	116	129	116	132							
7	119	124	124	116	129							
8	129	112	114	129	116							
9	124	119	119	124	119							
10	110	119	124	114								
11	124	112		116								
12	119			129								
13	124											
14	Групповые средние											
15	119,33	117,30	124,89	120,45	121,75							
16	Общая средняя											
17	120,56											
18	Групповые суммы квадратов отклонений											
19	394,67	186,10	256,89	328,73	251,50							
20	Общая сумма квадратов отклонений											
21	1722,32											
22	Общее число наблюдений			Число уровней фактора								
23	50			5								

Рисунок 3.13 – Вид листа MS Excel с исходными данными и расчетами для дисперсионного анализа

Результаты расчетов по дисперсионному анализу с помощью функций MS Excel оформим так же, как это делает *Пакет анализа* (рисунок 3.13).

Общая сумма квадратов отклонений уже нами рассчитана, поэтому в ячейке G4 просто поставим ссылку на ячейку A21.

Внутригрупповую (остаточную) сумму квадратов отклонений по формуле (3.6) рассчитаем в ячейке G3, сложив соответствующие значения для всех заводов (=СУММ(A19:E19)).

Межгрупповую (факторную) сумму квадратов отклонений найдем как разность значений в ячейках G4 и G3.

Введем в ячейки H2 и H3 число степеней свободы: для межгрупповой дисперсии – это  $k - 1 = 5 - 1 = 4$ , для внутригрупповой дисперсии –  $n - k = 50 - 5 = 45$ .

Рассчитаем межгрупповую и внутригрупповую дисперсии в ячейках I2 и I3, разделив соответствующие суммы квадратов отклонений на число степеней свободы.

Далее рассчитаем наблюдаемое значение критерия Фишера согласно формуле (3.7), разделив межгрупповую (факторную) дисперсию на внутригрупповую (остаточную) дисперсию. Таким образом,  $F_{\text{набл}} \approx 2,42$ .

Для расчета критической точки распределения Фишера в ячейке L2 используем функцию Excel ФРАСПОБР(0,05;H2;H3). Получим  $F_{\text{кр}} = 2,58$ . Поскольку  $F_{\text{набл}} < F_{\text{кр}}$ , можно считать несущественным влияние фактора и принять гипотезу о равенстве математических ожиданий генеральных совокупностей, соответствующих каждому заводу. Таким образом доказано, что расход сырья на производство исследуемого вида продукции не зависит от завода.

Рассчитаем также p-значение в ячейке K2 с помощью функции ФРАСПИ(J2;H2;H3). Полученный результат означает, что для всех уровней значимости, меньших либо равных 0,063, гипотеза о равенстве математических ожиданий уровней может быть принята. Поскольку  $\alpha = 0,05 < 0,063$ , влияние фактора признается несущественным.

Выборочный коэффициент детерминации  $\left( R^2 = \frac{Q_{\text{фактор}}}{Q_{\text{общ}}} = \frac{304,44}{1722,32} \approx 0,18 \right)$  рассчитан в ячейке F6. Он означает, что только 18% общей выборочной вариации расхода сырья связано с выбором завода.

Аналогичные результаты получим с помощью инструмента из *Пакета анализа*. Зададим команду *Сервис/Анализ данных...* и выберем *Однофакторный дисперсионный анализ*. Заполним диалоговое окно, как показано на рисунке 3.14.

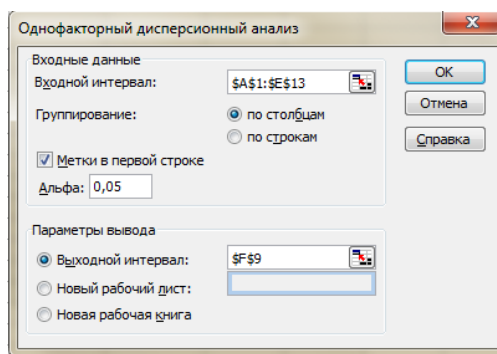


Рисунок 3.14 – Диалоговое окно «Однофакторный дисперсионный анализ»

Флажок *Метки в первой строке* поставлен потому, что входной интервал A1:E13 включает заголовки столбцов, и они будут использованы для формирования результата. Уровень значимости задан равным 0,05 (поле *Альфа*). Результат работы этого инструмента анализа представлен на рисунке 3.15.

	F	G	H	I	J	K	L
9	Однофакторный дисперсионный анализ						
10							
11	ИТОГИ						
12	Группы	Счет	Сумма	Среднее	Дисперсия		
13	Завод 1	12	1432	119,333333	35,878788		
14	Завод 2	10	1173	117,3	20,677778		
15	Завод 3	9	1124	124,888889	32,111111		
16	Завод 4	11	1325	120,454545	32,872727		
17	Завод 5	8	974	121,75	35,928571		
18							
19							
20	Дисперсионный анализ						
21	Источник вариации	SS	df	MS	F	P-Значение	F критическое
22	Между группами	304,437172	4	76,1092929	2,4155157	0,062563703	2,578739184
23	Внутри групп	1417,88283	45	31,5085073			
24							
25	Итого	1722,32	49				

Рисунок 3.15 – Результат работы инструмента  
Однофакторный дисперсионный анализ

В первой таблице результатов анализа показаны выборочные характеристики для каждого уровня фактора: количество наблюдений (счет), сумма значений, среднее и дисперсия.

Во второй таблице результатов показаны расчеты для дисперсионного анализа, аналогичные тем, что были ранее рассчитаны с помощью стандартных функций MS Excel. Вывод будет аналогичный.

**Пример 3.5.** Необходимо выяснить, влияет ли расстояние от центра города на степень заполняемости гостиниц. Пусть введены три уровня градации расстояния от центра: до 3 км; от 3 до 5 км; более 5 км. Данные по заполняемости гостиниц приведены в таблице 3.15.

Таблица 3.15 – Заполняемость гостиниц при различных расстояниях  
от центра города, %

Расстояние	Заполняемость гостиниц					
	1-я	2-я	3-я	4-я	5-я	6-я
До 3 км	92	98	89	97	90	94
От 3 до 5 км	90	86	84	91	83	82
Более 5 км	87	79	74	85	73	77

#### Решение

Введем исходные данные на лист Excel, как показано на рисунке 3.16. Зададим команду *Сервис/Анализ данных* и выберем инструмент *Однофакторный дисперсионный анализ*. Заполним диалоговое окно аналогично предыдущему примеру. Результаты анализа показаны на рисунке 3.16.

	A	B	C	D	E	F	G
1	Заполняемость гостиниц						
2	до 3 км	от 3 до 5 км	свыше 5 км				
3	92	90	87				
4	98	86	79				
5	89	84	74				
6	97	91	85				
7	90	83	73				
8	94	82	77				
9							
10	Однофакторный дисперсионный анализ						
11							
12	ИТОГИ						
13	Группы	Счет	Сумма	Среднее	Дисперсия		
14	до 3 км	6	560	93,33333	13,4666667		
15	от 3 до 5 км	6	516	86	14		
16	свыше 5 км	6	475	79,16667	32,9666667		
17							
18							
19	Дисперсионный анализ						
20	Источник вариации	SS	df	MS	F	P-Значение	F критическое
21	Между группами	602,333333	2	301,1667	14,9503585	0,000268401	3,682320344
22	Внутри групп	302,166667	15	20,14444			
23							
24	Итого	904,5	17				

Рисунок 3.16 – Результаты дисперсионного анализа влияния фактора расстояния от центра на заполняемость гостиниц

Поскольку  $F_{\text{факт}} = 14,95 > F_{\text{кр}} = 3,68$ , влияние фактора расстояния от центра города должно быть признано существенным, а гипотеза о равенстве математических ожиданий уровней отклоняется. О том же свидетельствует и тот факт, что  $p$ -значение значительно меньше уровня значимости 0,05.

Рассчитаем выборочный коэффициент детерминации:  $R^2 = \frac{602,33}{904,5} \approx 0,67$ . Таким образом, 67% общей выборочной вариации связано с действием фактора расстояния.

### Задания для самостоятельной работы

1. Определите, влияет ли фактор образования на уровень заработной платы работников гостиницы на основании данных, приведенных в таблице 3.16. Используйте уровень значимости 10%. Определите, можно ли достоверно считать, что фактор образования имеет влияние при уровне значимости 5%.

Таблица 3.16 – Данные наблюдений заработной платы сотрудников гостиницы

Образование	Заработная плата сотрудников, тыс. р.					
	1-й месяц	2-й месяц	3-й месяц	4-й месяц	5-й месяц	6-й месяц
Высшее	3 200	3 000	2 600	2 000	1 900	1 900
Среднее специальное	2 600	2 000	2 000	1 900	1 800	1 700
Среднее	2 000	2 000	1 900	1 800	1 700	1 700

2. На химическом заводе разработаны два новых варианта технологического процесса. Чтобы оценить, как изменится дневная производительность при переходе на работу по новым технологиям, завод в течение 10 дней работал по каждому варианту, включая существующий. Дневная производительность завода (в условных единицах) представлена в таблице 3.17. При уровне значимости 0,01 выясните, как зависит дневная производительность от технологического процесса. Оцените степень этой зависимости.

Таблица 3.17 – Данные о дневной производительности по различным технологиям, усл. ед.

Дни работы	Дневная производительность		
	Существующая технология	Вариант 1	Вариант 2
1-й	119	53	76
2-й	26	128	149
3-й	96	176	97
4-й	174	106	77
5-й	71	96	114
6-й	101	172	197
7-й	135	81	179
8-й	171	163	178
9-й	148	73	90
10-й	20	108	109

3. В исследовании изучалась эффективность трех рекламных роликов А, Б, В. Для оценки рекламы по девятибалльной шкале выбрали 10 потребителей. Полученные данные приведены в таблице 3.18. При уровне значимости 0,05 выясните, какой ролик можно считать более эффективным.

Таблица 3.18 – Данные об оценке рекламных роликов потребителями (в баллах)

Рекламные ролики	Потребители									
	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	10-й
А	4	5	3	4	3	7	4	3	5	5
Б	7	5	6	5	4	6	5	5	4	4
В	8	7	7	6	8	7	6	8	7	6

4. В таблице 3.19 приведены данные оценки влияния брэндов трех торговых сетей на их товарооборот шести респондентов. Определите, брэнд какой торговой сети больше влияет на товарооборот, и можно ли считать это влияние значимым при  $\alpha = 0,05$ .

Таблица 3.19 – Данные об оценке влияния брэндов торговых сетей на товарооборот (в баллах)

Торговые сети	Респонденты					
	1-й	2-й	3-й	4-й	5-й	6-й
Первая	6	8	10	7	6	7
Вторая	8	9	7	10	9	8
Третья	10	9	8	10	10	10

5. В таблице 3.20 приведены данные о влиянии технологий организации торговли на увеличение объема продаж в различных торговых организациях, где они были внедрены. Определите, влияет ли фактор технологии на рост объемов продаж при уровне значимости 0,05.

Таблица 3.20 – Увеличение объемов продаж в различных торговых организациях, %

Технологии торговли	Увеличение объема продаж			
	1-я организация	2-я организация	3-я организация	4-я организация
Первая	7,7	7,9	8,2	7,7
Вторая	9,4	9,1	8,9	8,8
Третья	9,1	8,9	7,9	8,6
Четвертая	9,3	9,3	8,9	9,1

6. Было исследовано влияние различных доз минеральных удобрений на урожайность озимой ржи. Результаты для различных участков посевных площадей приведены в таблице 3.21. Проведите дисперсионный анализ при уровне значимости 0,01. Определите, можно ли считать влияние размера вносимой дозы удобрения статистически значимым.

Таблица 3.21 – Данные наблюдений за урожайностью озимой ржи при различных дозах минеральных удобрений

Дозы удобрений, кг/га	Данные об урожайности, ц/га			
	1-й участок	2-й участок	3-й участок	4-й участок
15	8,0	8,4	9,0	8,6
20	8,2	9,0	10,0	10,0
25	11,0	13,0	12,0	–
30	7,5	8,5	–	–

7. В таблице 3.22 приведены данные о влиянии образования рабочих (неполное среднее, среднее, профессионально-техническое) на их заработную плату. В каждой группе было опрошено по 4 чел. Проведите дисперсионный анализ при уровне значимости 0,05. Определите, можно ли считать влияние фактора образования на заработную плату рабочих статистически достоверным.

Таблица 3.22 – Данные для дисперсионного анализа

Образование	Заработная плата рабочих, усл. ед.			
	1-й рабочий	2-й рабочий	3-й рабочий	4-й рабочий
Неполное среднее	175	183	191	120
Среднее	180	210	165	150
Профессионально-техническое	190	220	210	290

8. Данные о влиянии маркетинговых мероприятий на увеличение объемов продажи продукции шести видов приведены в таблице 3.23. С помощью дисперсионного анализа выясните, влияет ли вид применяемых маркетинговых мероприятий на показатели продажи продукции. Используйте уровень значимости 10%.

Таблица 3.23 – Данные наблюдений об увеличении продажи продукции, %

Виды маркетинговых мероприятий	Вид продукции					
	1-й	2-й	3-й	4-й	5-й	6-й
А	21	20	19	18	20	21
Б	20	22	21	19	18	19
В	17	16	18	19	17	16
Г	11	10	11	13	11	10

### 3.3. Ковариация и корреляция

Одной из основных задач экономического анализа является выявление наличия взаимосвязей между различными экономическими показателями и определение силы этих связей. Например, между доходом и потреблением, между спросом на конкретный товар и его ценой имеется взаимосвязь, причем обычно она носит не функциональный (строго однозначный), а статистический характер.

*Статистической* называют зависимость между случайными величинами, при которой изменение одной из них влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. В этом случае статистическую зависимость называют *корреляционной*.

Приведем пример статистической взаимосвязи между двумя случайными величинами  $X$  и  $Y$ . Пусть  $X$  – количество удобрений, а  $Y$  – урожай зерна. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, ( $Y$  не является функцией от  $X$ ). Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, *средний* урожай является функцией от количества удобрений, т. е.  $Y$  связан с  $X$  статистической зависимостью.

Корреляционный анализ позволяет на основе выборочных данных оценить наличие, направленность и силу статистической взаимосвязи.

По направлению связи бывают *прямыми*, когда зависимая переменная растет с увеличением факторного признака, и *обратными*, при которых рост факторного признака сопровождается уменьшением зависимой переменной. Такие связи также можно назвать положительными и отрицательными.

Пусть для двух показателей  $X$  и  $Y$  (случайных величин) имеется выборка связанных пар наблюдений  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , где  $n$  – число наблюдений.

*Выборочной ковариацией*  $\overline{\text{cov}}(X; Y)$  называется величина

$$\overline{\text{cov}}(X; Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad (3.9)$$

где  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  и  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  – средние значения выборочных данных для величин  $X$  и  $Y$  соответственно.

Ковариация является мерой взаимосвязи случайных величин. Если ковариация равна нулю, то взаимосвязь величин отсутствует. Если  $\overline{\text{cov}}(X, Y) > 0$ , то существует прямая взаимосвязь, а если  $\overline{\text{cov}}(X, Y) < 0$  – обратная взаимосвязь случайных величин. Однако эта характеристика обладает рядом существенных недостатков. Во-первых, ее значение зависит от единиц измерения исследуемых случайных величин. Во-вторых, она не позволяет оценить силу зависимости между ними. Для устранения данных недостатков вводится относительная мера взаимосвязи (безразмерная величина) – коэффициент корреляции. Существует несколько типов коэффициентов корреляции. Мы рассмотрим только один: коэффициент линейной корреляции (коэффициент Пирсона), который характеризует степень линейной взаимосвязи двух случайных величин.

*Выборочным коэффициентом линейной корреляции* ( $r_{XY}$ ) случайных величин  $X$  и  $Y$  называется величина, определяемая по формуле

$$r_{XY} = \frac{\overline{\text{cov}(X;Y)}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}. \quad (3.10)$$

Коэффициент линейной корреляции всегда удовлетворяет соотношению:  $-1 \leq r_{XY} \leq 1$ .

Если  $r_{XY} = 0$ , то линейная взаимосвязь между случайными величинами отсутствует. Это может означать, что данные случайные величины независимы, либо между ними существует нелинейная взаимосвязь (например, показательная, логарифмическая и др.).

Если  $0 < r_{XY} < 1$ , то между  $X$  и  $Y$  существует прямая линейная зависимость. Это означает, что увеличение одного признака ведет к увеличению другого. Например, при увеличении температуры возрастает давление газа.

Если  $-1 < r_{XY} < 0$ , то между  $X$  и  $Y$  имеется обратная линейная зависимость. Обратная зависимость означает, что увеличение одного признака ведет к уменьшению другого. Например, связь между температурой воздуха и количеством топлива, расходуемого на обогрев помещения.

Если  $r_{XY} = \pm 1$ , то между  $X$  и  $Y$  существует линейная функциональная зависимость.

Степень линейной зависимости можно качественно оценить с помощью шкалы Чеддока (таблица 3.24).

Таблица 3.24 – Шкала Чеддока

Значение коэффициента корреляции (по модулю)	Теснота связи
0,1–0,3	Слабая
0,3–0,5	Умеренная
0,5–0,7	Заметная
0,7–0,9	Высокая
0,9–0,99	Весьма высокая

В MS Excel для вычисления парных коэффициентов линейной корреляции используется функция КОРРЕЛ(*массив1*; *массив2*), где *массив1* и *массив2* – это диапазоны ячеек, содержащих выборочные значения первой и второй случайной величины.

При исследовании связи между несколькими случайными величинами находят выборочные коэффициенты корреляции между парами всех исследуемых величин и строят корреляционную матрицу.

*Корреляционная матрица* – это квадратная таблица, в которой на пересечении строки  $i$  и столбца  $j$  находится коэффициент корреляции ( $r_{ij}$ ) между случайными величинами  $X_i$  и  $X_j$ . Эта матрица является симметричной, поэтому часто указывается только половина таблицы (например, под главной диагональю). По диагонали стоят единицы, так как каждая величина полностью коррелирует сама с собой.

Например, корреляционная матрица для трех показателей  $X$ ,  $Y$ ,  $Z$  имеет следующий вид:

$$\begin{pmatrix} 1 & r_{XY} & r_{XZ} \\ r_{YX} & 1 & r_{YZ} \\ r_{ZX} & r_{ZY} & 1 \end{pmatrix}.$$

В MS Excel для расчета корреляционной матрицы используется инструмент *Корреляция* из *Пакета анализа*.

Выборочный коэффициент корреляции ( $r_{XY}$ ) является оценкой коэффициента корреляции генеральной совокупности ( $\rho_{XY}$ ). Допустим, выборочный коэффициент корреляции, рассчитанный по формуле (3.10), оказался отличным от нуля. Так как выборка отобрана случайно, то еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. Возникает необходимость проверить гипотезу о значимости (существенности) выборочного коэффициента корреляции или, что то же самое, о равенстве нулю коэффициента корреляции генеральной совокупности.

Выдвигается нулевая гипотеза о равенстве нулю коэффициента корреляции генеральной совокупности  $H_0: \rho_{XY} = 0$  при альтернативной гипотезе  $H_1: \rho_{XY} \neq 0$ . Если гипотеза  $H_0$  будет отвергнута, то выборочный коэффициент корреляции значим, а между величинами  $X$  и  $Y$  существует линейная взаимосвязь. Если же гипотеза  $H_0$  будет принята, то выборочный коэффициент корреляции незначим, а линейная связь между  $X$  и  $Y$  отсутствует.

При проверке нулевой гипотезы используется  $t$ -статистика, которая имеет распределение Стьюдента с степенями свободы  $n - 2$ :

$$t = \frac{|r_{XY}|}{\sqrt{1 - r_{XY}^2}} \sqrt{n - 2}.$$

По выборке находится наблюдаемое значение статистики ( $t_{набл.}$ ). Для заданного уровня значимости  $\alpha$  определяется критическая точка распределения Стьюдента:  $t_{кр} = t(\alpha; n - 2)$ . В MS Excel для определения этого значения используется функция СТЬЮДРАСПОБР(*вероятность; степени\_свободы*).

**Пример 3.6.** Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков (таблица 3.25). Необходимо определить, существует ли линейная взаимосвязь между этими случайными величинами и оценить степень этой взаимосвязи.

Таблица 3.25 – Фактические данные о количестве посетителей музея и парка, чел.

Число ясных дней	Количество посетителей	
	музея	парка
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

#### Решение

Введем исходные данные, как показано на рисунке 3.17. Зададим команду *Сервис/Анализ данных...* и выберем инструмент *Корреляция*. Заполним диалоговое окно «Корреляция» как показано на рисунке 3.18.

	А	В	С	Д
1	Число ясных дней	Количество посетителей музея	Количество посетителей парка	
2	8	495	132	
3	14	503	348	
4	20	380	643	
5	25	305	865	
6	20	348	743	
7	15	465	541	
8				
9		Число ясных дней	Количество посетителей музея	Количество посетителей парка
10	Число ясных дней	1		
11	Количество посетителей музея	-0,921854339	1	
12	Количество посетителей парка	0,974575588	-0,919375244	1

Рисунок 3.17 – Данные примера 3.5. и корреляционная матрица, сформированная *Пакетом анализа*



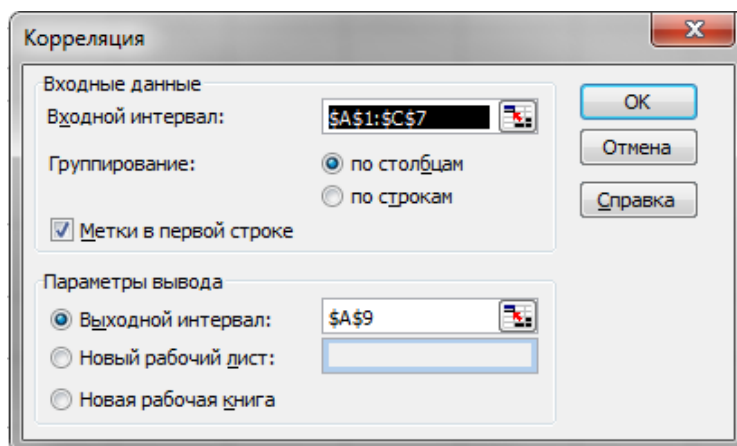


Рисунок 3.18 – Пример заполнения диалогового окна «Корреляция»

*Входной интервал* охватывает все фактические данные, причем каждой случайной величине отведен отдельный столбец, на что указывает переключатель *Группирование*. Первая строка, содержащая заголовки столбцов, также вошла в диапазон входного интервала, поэтому мы установили флажок *Метки в первой строке*. Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, чтобы корреляционная матрица расположилась на текущем листе. В соответствующем поле указан адрес левой верхней ячейки диапазона, в который будет выведена эта матрица.

Таким образом, коэффициент линейной корреляции между числом ясных дней и количеством посетителей музея ( $r_{12}$ ) равен  $-0,92$ , что говорит о весьма высокой степени обратной линейной связи (по шкале Чеддока); между числом ясных дней и количеством посетителей парка ( $r_{13}$ ) данный коэффициент равен  $0,97$ , что означает практически функциональную линейную прямую зависимость этих величин.

Коэффициент корреляции между количеством посетителей музея и парка ( $r_{23}$ ) равен  $-0,92$ , т. е. имеется весьма высокая обратная линейная зависимость.

Аналогичную матрицу легко получить также с помощью функции КОРРЕЛ(), задавая ее последовательно для каждой пары диапазонов случайных величин (рисунок 3.19).

	А	В	С
13	Корреляционная матрица, рассчитанная с помощью функции КОРРЕЛ()		
14	1		
15	=КОРРЕЛ(A2:A7;B2:B7)	1	
16	=КОРРЕЛ(A2:A7;C2:C7)	=КОРРЕЛ(B2:B7;C2:C7)	1

Рисунок 3.19 – Расчет корреляционной матрицы с помощью функции КОРРЕЛ()

Проверим каждый коэффициент корреляции на значимость. Для этого найдем наблюдаемые значения  $t$ -статистики по формуле (3.11) (учитываем, что  $n = 6$ ). Оформим фрагмент листа, как показано на рисунке 3.20, и в ячейку A20 введем следующую формулу MS Excel:

$$=ABS(B11)/КОРЕНЬ(1-B11^2)*КОРЕНЬ(6-2).$$

Эту формулу затем скопируем по таблице вниз и вправо методом автозаполнения («протянем»).

Для расчета критического значения  $t$ -статистики при уровне значимости  $0,05$  и числе степеней свободы  $n - 2 = 4$  введем в ячейку A23 формулу =СТЮДРАСПОБР(0,05;4).

Результаты расчетов показаны на рисунке 3.20.

17	Наблюдаемые значения t-критерия			
18		Число ясных дней	Количество посетителей музея	Количество посетителей парка
19	Число ясных дней			
20	Количество посетителей музея	4,757509125		
21	Количество посетителей парка	8,699277522	4,674214639	
22	Критическое значение t-критерия			
23	2,776445105			

Рисунок 3.20 – Наблюдаемые и критическое значения *t*-статистики

Поскольку все наблюдаемые значения статистики больше критического, следует отвергнуть нулевые гипотезы о равенстве нулю каждого коэффициента корреляции генеральной совокупности. Все рассчитанные выборочные коэффициенты корреляции признаются значимыми.

### Задания для самостоятельной работы

1. Исследуется связь между расходами дилеров некоторой компании на рекламу продукции (*X*, тыс. усл. ед.) и объемами ее продаж (*Y*, тыс. усл. ед.). Выясните, существует ли линейная корреляционная зависимость между *X* и *Y*, используя сведения, которые приведены в таблице 3.26, по 15 случайно отобраным дилерам.

Таблица 3.26 – Данные о расходах на рекламу и объемах продаж продукции, усл. ед.

Дилеры	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	10-й	11-й	12-й	13-й	14-й	15-й
<i>X</i>	7	36	5	69	45	21	8	17	56	23	16	9	25	31	10
<i>Y</i>	215	458	165	369	698	560	752	387	789	590	960	625	800	670	456

2. Показатели уровня образования, уровня преступности, а также отношение количества безработных к числу вакансий в областях Беларуси в 2002 г. приведены в таблице 3.27 (уровень образования рассчитывался как численность лиц с высшим и средним специальным образованием на 1000 жителей области, уровень преступности – как число совершенных преступлений на 100 тыс. жителей области). Установите наличие линейной взаимосвязи между указанными показателями.

Таблица 3.27 – Данные о социальном состоянии областей

Области	Уровень образования	Отношение количества безработных к числу вакансий	Уровень преступности
Минская	788	10,8	791
Гомельская	735	22,3	908
Могилевская	779	52,9	804
Брестская	740	10,4	685
Витебская	763	5,4	936

3. Данные о прибыли, оборотных средствах, стоимости основных фондов шести предприятий приведены в таблице 3.28. Установите наличие линейной взаимосвязи между указанными показателями.

Таблица 3.28 – Показатели работы предприятий, усл. ед.

Предприятия	Прибыль	Величина оборотных средств	Стоимость основных фондов
1-е	188	129	791
2-е	78	64	908
3-е	93	69	804
4-е	152	87	685
5-е	55	53	936
6-е	161	140	750

4. Исследуется зависимость заработной платы специалистов определенной профессии от возраста, количества лет обучения (колледж, вуз (неполное высшее образование), курсы повышения квалификации) и стажа работы по данной специальности. Случайным образом выбраны 12 специалистов. Результаты исследования приведены в таблице 3.29. Установите наличие линейной взаимосвязи между указанными показателями.

Таблица 3.29 – Данные о специалистах исследуемой профессии

Номер специалиста по списку	Заработная плата, усл. ед.	Количество лет обучения	Возраст, лет	Стаж работы по специальности, лет
1	225	5	28	10
2	197	3	26	8
3	198	3	30	14
4	210	7	35	15
5	309	9	41	16
6	402	13	45	18
7	220	6	27	3
8	207	7	35	13
9	207	4	30	10
10	100	0	23	2
11	227	7	29	7
12	214	8	33	11

5. Определите наличие линейной зависимости объема потребления по домохозяйству ( $Y$ ) в зависимости от располагаемого дохода ( $X$ ) его членов по данным 12 наблюдений, приведенных в таблице 3.30.

Таблица 3.30 – Данные о доходах и объемах потребления домохозяйств, усл. ед.

Номер наблюдения	1	2	3	4	5	6	7	8	9	10	11	12
$X$	107	109	110	120	115	122	123	136	140	145	150	165
$Y$	102	105	109	115	112	120	119	125	136	129	141	136

6. По данным работы 20 предприятий (таблица 3.31) установите существование линейной связи между показателями фондоотдачи и удельного веса продукции высшей категории качества. Определите силу этой связи (при ее наличии).

Таблица 3.31 – Данные о работе предприятий

Предприятия	Фондоотдача ( $Y$ ), усл. ед.	Удельный вес продукции высшей категории качества ( $X$ ), %	Предприятия	Фондоотдача ( $Y$ ), усл. ед.	Удельный вес продукции высшей категории качества ( $X$ ), %
1-е	147	34,08	11-е	191	40,89
2-е	125	35,89	12-е	183	47,63
3-е	182	36,93	13-е	181	45,63
4-е	145	32,31	14-е	151	43,07
5-е	175	34,91	15-е	154	44,16
6-е	137	30,2	16-е	198	35,22
7-е	161	31,23	17-е	169	38,83
8-е	193	48,13	18-е	168	39,69
9-е	168	30,08	19-е	176	39,39
10-е	166	42,86	20-е	173	43,17

7. Проведите корреляционный анализ зависимости средней заработной платы от средней производительности на различных предприятиях, используя данные таблицы 3.32.

Таблица 3.32 – Данные о заработной плате и производительности труда на предприятиях, усл. ед.

Предприятия	Средняя производительность	Средняя заработная плата
1-е	9 320	3 320
2-е	8 630	3 640
3-е	8 050	3 900
4-е	8 320	4 120
5-е	8 600	4 090
6-е	9 120	4 200
7-е	9 540	4 380
8-е	9 730	4 500
9-е	10 120	4 610
10-е	10 740	4 800
11-е	11 200	5 000

8. Определите наличие и силу линейной зависимости между количеством покупателей и объемом продаж торгового предприятия по данным наблюдений за 20 дней (таблица 3.33).

Таблица 3.33 – Данные наблюдений о работе торгового предприятия за 20 дней

Номер наблюдения	Количество покупателей (X), чел.	Объем продаж (Y), усл. ед.	Номер наблюдения	Количество покупателей (X), чел.	Объем продаж (Y), усл. ед.
1	188	49 980	11	192	50 870
2	163	39 710	12	195	51 890
3	140	31 380	13	140	47 000
4	193	49 010	14	205	55 000
5	147	48 170	15	195	56 000
6	155	24 110	16	151	49 280
7	190	50 560	17	140	33 330
8	133	26 130	18	113	38 960
9	122	24 730	19	174	48 000
10	105	36 480	20	179	49 000

### 3.4. Регрессионный анализ

Если корреляционный анализ позволяет оценить наличие и силу статистической взаимосвязи случайных величин, то целью регрессионного анализа является установление формы этой зависимости. Такая форма определяется в виде некоторой функции зависимой величины  $Y$  от независимых величин  $X_1, X_2, \dots, X_k$  (факторов), которая называется *уравнением регрессии*.

Если исследуется зависимость случайной величины  $Y$  от одного фактора  $X$ , то модель называется *однофакторной* (или *парной*). Если же число независимых случайных величин два и больше ( $k \geq 2$ ), то регрессионная модель называется *многофакторной* (или *множественной*). Различают также линейную и нелинейную регрессии. Уравнение линейной однофакторной регрессии имеет вид

$$\tilde{Y} = a_1 X + a_0.$$

Уравнение множественной линейной регрессии имеет вид

$$\tilde{Y} = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_k X_k,$$

где  $a_0, a_1, \dots, a_k$  – параметры регрессии, которые необходимо определить по выборочным данным.

В случае, если зависимость  $Y$  от  $X_1, X_2, \dots, X_j$  выражается любой нелинейной функцией, то регрессия называется *нелинейной*. Примером является экспоненциальная (показательная) многофакторная модель:

$$\tilde{Y} = a \cdot b_1^{X_1} \cdot b_2^{X_2} \cdot \dots \cdot b_n^{X_k}.$$

Ограничимся рассмотрением многофакторной линейной регрессии. Исходной информацией для построения уравнения регрессии является выборка, состоящая из наборов наблюдений  $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$ , где  $i = \overline{1, n}$ . На основе этих наблюдений необходимо подобрать параметры регрессии таким образом, чтобы уравнение регрессии наилучшим образом приближалось к фактическим данным.

Определение параметров регрессии выполняется на основе *метода наименьших квадратов*. Суть этого метода состоит в минимизации суммы квадратов отклонений фактических значений зависимой переменной от теоретических, т. е. найденных по уравнению регрессии:

$$\sum_{i=1}^n [y_i - \tilde{y}(x_{1i}, x_{2i}, \dots, x_{ki})]^2 \rightarrow \min. \quad (3.12)$$

В пакете MS Excel для определения параметров линейной регрессии можно использовать функцию ЛИНЕЙН(), а также процедуру *Регрессия* из *Пакета анализа*.

Качество однофакторной регрессионной модели оценивается с помощью *коэффициента детерминации*:

$$R^2 = 1 - \frac{\sum_{i=1}^n (\tilde{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.13)$$

где  $y_i$  – фактическое значение зависимой величины в  $i$ -м наблюдении;

$\tilde{y}_i = \tilde{y}(x_{1i}, x_{2i}, \dots, x_{ki})$  – значение зависимой переменной, определяемое по уравнению регрессии;

$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  – среднее арифметическое фактических значений зависимой переменной.

Коэффициент детерминации может принимать значения от 0 до 1. Чем больше коэффициент детерминации, тем более точной считается модель. В случае, когда  $R^2 < 0,6$ , считают, что точность приближения недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейности и т. д.). Если коэффициент детерминации  $R^2 \geq 0,9$ , то регрессия считается достаточно точной для того, чтобы использовать ее для практических расчетов. Если значения зависимой переменной, найденные по уравнению регрессии, совпадают с фактическими данными, то  $R^2 = 1$ .

Поскольку коэффициент детерминации рассчитывается по выборке, его значение является случайным числом. Необходимо доказать значимость этого коэффициента, т. е. его соответствие истинной зависимости между  $Y$  и исследуемыми факторами. Доказательство значимости регрессионной модели выполняется на основании *критерия Фишера* в следующей последовательности:

1. Выдвигается гипотеза  $H_0$ :  $R^2 = 0$ , т. е. линейная связь между исследуемыми показателями отсутствует.

2. Рассчитывается выборочное значение  $F$ -критерия по формуле

$$F = \frac{R^2}{(1 - R^2)} \cdot \frac{n - k - 1}{k},$$

где  $n$  – число наблюдений,

$k$  – число факторов (независимых переменных  $X_i$ ).

3. По таблице или с помощью функции `ФРАСПОБР()` вычисляется критическое значение критерия Фишера ( $F_{\text{ед}}$ ) с уровнем значимости  $\alpha$  и числом степеней свободы  $k$  и  $(n - k - 1)$ .

4. Если  $F > F_{\text{ед}}$ , то нулевая гипотеза отклоняется и значимость регрессии признается.

5. Пакет Excel рассчитывает также  $p$ -значение для данной гипотезы (в поле *Значимость F* из *Пакета анализа*). Если  $p$ -значение меньше уровня значимости ( $p < \alpha$ ), то нулевая гипотеза отклоняется (модель значима).

Параметры регрессионной модели, которые дает метод наименьших квадратов для данной выборки, по сути дела являются только оценками соответствующих математических ожиданий параметров. Процедура *Регрессия* из *Пакета анализа* рассчитывает также стандартную ошибку для каждого параметра ( $m_i = s_i / \sqrt{n}$ , где  $s_i$  – стандартное отклонение параметра).

Чтобы доказать, что данный фактор  $X_i$  действительно имеет влияние на зависимую величину  $Y$ , на основании *критерия Стьюдента* проверяется достоверность отличия от нуля параметра  $a_i$  ( $i = 1, 2, \dots, k$ ). Эта проверка осуществляется следующим образом:

1. Выдвигается гипотеза  $H_0$ : «Параметр регрессионной модели  $a_i = 0$ ».

2. Рассчитывается наблюдаемое значение  $t$ -критерия:

$$t = \frac{a_i}{m_i}. \quad (3.15)$$

3. По таблице или с помощью функции `СТЮДРАСПОБР()` определяется критическое значение критерия Стьюдента ( $t_{\text{кр}}$ ) для уровня значимости  $\alpha$  и числа степеней свободы  $n - 2$ .

4. Если  $t > t_{\text{ед}}$ , то нулевая гипотеза должна быть отвергнута,  $a_i \neq 0$  и фактор является значимым.

5. В пакете Excel рассчитывается также  $p$ -значение для нулевой гипотезы. Если  $p < \alpha$ , то нулевая гипотеза отклоняется и соответствующий параметр считается достоверно отличным от нуля.

В случае, если в результате расчетов будет принято, что  $a_i = 0$ , соответствующий фактор нужно исключить из модели.

Аналогично можно проверить гипотезу о равенстве нулю свободного члена уравнения регрессии  $a_0$ .

Целью построения регрессионных моделей является прогнозирование значения зависимой переменной  $Y$  для заданных значений независимых переменных  $X_1, X_2, \dots, X_k$ . Для расчета прогноза нужно просто подставить эти значения в уравнение регрессии. Необходимо помнить, что при этом получается точечный прогноз, относительно которого могут быть отклонения в пределах некоторого доверительного интервала.

**Пример 3.7.** Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания ( $Y$ ) в зависимости от содержания в воздухе двуокиси углерода ( $X_1$ ) и степени запыленности ( $X_2$ ). В таблице 3.34 приведены данные наблюдений в течение 28 месяцев. Предсказать уровень заболеваемости при содержании двуокиси углерода, равной 0,7, и запыленности – 1,5.

Таблица 3.34 – Данные об уровне заболеваемости

Месяц	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	10-й	11-й	12-й	13-й	14-й
$X_1$	1	1	1,1	1,1	1,1	1,1	1	1	1,2	1,2	0,6	0,6	0,7	0,7
$X_2$	1,3	1,3	1,4	1,4	1,5	1,5	1,4	1,5	1,6	1,7	1	1	1,1	1,15
$Y$	1 160	1 155	1 158	1 157	1 160	1 161	1 157	1 159	1 256	1 260	1 040	1 039	1 039	1 040

Продолжение таблицы 3.34

Месяц	15-й	16-й	17-й	18-й	19-й	20-й	21-й	22-й	23-й	24-й	25-й	26-й	27-й	28-й
$X_1$	0,75	0,7	0,7	0,7	0,8	0,8	0,78	0,8	0,78	0,78	0,8	0,8	0,75	0,78
$X_2$	1,2	1,2	1,3	1,3	1,4	1,4	1,5	1,5	1,5	1,6	1,7	1,8	1,8	1,9
$Y$	1 040	1 039	1 040	1 039	1 140	1 138	1 240	1 239	1 241	1 240	1 239	1 239	1 240	1 238

#### Решение

Введем исходные данные, расположив каждую случайную величину в отдельном столбце (на рисунке 3.21 показаны первые 14 строк исходных данных).

	А	В	С
1	Содержание CO <sub>2</sub> ( $X_1$ )	Запыленность ( $X_2$ )	Уровень заболеваемости ( $Y$ )
2	1	1,3	1160
3	1	1,3	1155
4	1,1	1,4	1158
5	1,1	1,4	1157
6	1,1	1,5	1160
7	1,1	1,5	1161
8	1	1,4	1157
9	1	1,5	1159
10	1,2	1,6	1256
11	1,2	1,7	1260
12	0,6	1	1040
13	0,6	1	1039
14	0,7	1,1	1039
15	0,7	1,15	1040

Рисунок 3.21 – Исходные данные для регрессионной модели

Для начала попробуем оценить наличие и силу линейных связей между зависимой величиной  $Y$  и каждым из факторов  $X_1$  и  $X_2$ . Для этого рассчитаем коэффициенты линейной корреляции, введя в свободные ячейки формулы =КОРРЕЛ(C2:C29;A2:A29) и =КОРРЕЛ(C2:C29;B2:B29). Получим:  $r_{YX_1}^1 = 0,475$  и  $r_{YX_2}^2 = 0,875$ . Таким образом, связь между заболеваемостью и содержанием CO<sub>2</sub> по шкале Чеддока является умеренной, а между заболеваемостью и запыленностью – высокой. Это означает, что фактор запыленности нужно включить в линейную модель, а фактор содержания CO<sub>2</sub> остается под вопросом.

Все же попытаемся построить двухфакторную регрессионную модель вида  $\tilde{y} = a_0 + a_1X_1 + a_2X_2$ . Выполним команду *Сервис/Анализ данных...* и выберем инструмент *Регрессия*. Заполним диалоговое окно, как показано на рисунке 3.22.

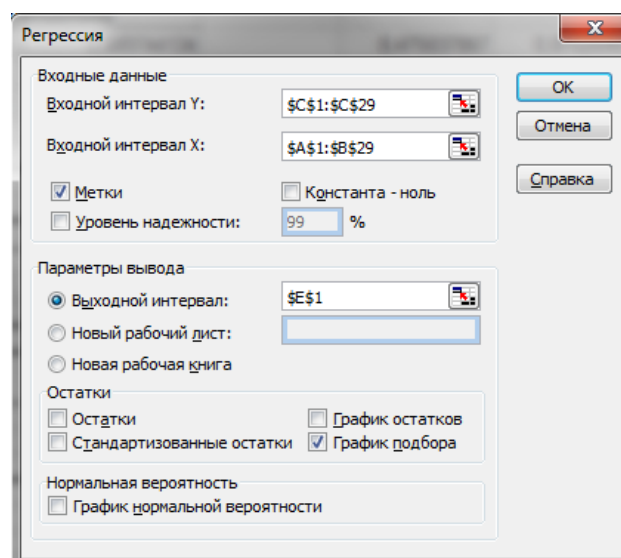


Рисунок 3.22 – Пример заполнения диалогового окна «Регрессия»

Элементы этого окна имеют следующие значения:

- Поле *Входной интервал Y* содержит диапазон значений зависимой переменной (заболеваемость органов дыхания).
- Поле *Входной интервал X* содержит диапазон из двух столбцов, в которых расположены значения независимых переменных (факторов) – содержание CO<sub>2</sub> и запыленность.
- Флажок *Метки* установлен, поскольку первые элементы отмеченных диапазонов содержат названия переменных (столбцов).
- Флажок *Константа-ноль* указывает на отсутствие свободного члена в уравнении регрессии (в нашем случае он не установлен, поскольку уравнение регрессии рассчитывается обычным образом).
- Флажок *Уровень надежности* устанавливается в том случае, когда необходимо изменить уровень значимости, заданный по умолчанию ( $\alpha = 0,05$ ).
- Переключатель *Параметры вывода* установим в положение *Выходной интервал*, чтобы поместить результат работы надстройки на текущем листе. В соответствующем поле укажем левый верхний угол диапазона, в котором будет выведен результат. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможности наложения выходного диапазона на другие данные.
- Флажок *График подбора* установлен для того, чтобы можно было визуально проверить соответствие регрессионной модели фактическим данным.

Результаты работы инструмента *Регрессия* из *Пакета анализа* показаны на рисунке 3.23.

ВЫВОД ИТОГОВ					
<i>Регрессионная статистика</i>					
Множественный R	0,889744134				
R-квадрат	0,791644624				
Нормированный R-квадрат	0,774976194				
Стандартная ошибка	39,3331654				
Наблюдения	28				
<i>Дисперсионный анализ</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	2	146954,6596	73477,32982	47,4936523	3,05534E-09
Остаток	25	38677,44751	1547,0979		
Итого	27	185632,1071			
	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>
Y-пересечение	672,4988584	51,26494841	13,11810271	1,0396E-12	566,9167217
Содержание CO <sub>2</sub> (X1)	79,69364308	45,42450365	1,754419678	0,09161008	-13,85987266
Запыленность (X2)	288,8816344	35,05502985	8,240804118	1,3626E-08	216,6844495
	<i>t критическое</i>	2,055529418		<i>F критическое</i>	3,385189962

Рисунок 3.23 – Результаты регрессионного анализа

Значения коэффициентов регрессии находятся в столбце *Коэффициенты* и соответствуют:

- *Y-пересечение* –  $a_0$ ;
- *содержание CO<sub>2</sub>* –  $a_1$ ;
- *запыленность* –  $a_2$ .

Таким образом, получаем следующее уравнение регрессии:

$$\tilde{Y} = 672,5 + 79,7 \cdot X_1 + 288,9 \cdot X_2.$$

Для каждого коэффициента рассчитана также стандартная ошибка и выборочное значение  $t$ -статистики (отношение оценки параметра к ее стандартной ошибке). Для оценки достоверности отличия каждого параметра от нуля найдем критическое значение критерия Стьюдента для уровня значимости 0,05. Введем в произвольную ячейку формулу =СТЮДРАСПОБР(0,05;26), где  $26 = n - 2$ . Расчет по этой формуле дает значение 2,056. Сравнивая это значение с  $t$ -статистикой для каждого параметра, убеждаемся, что для  $a_0$  и  $a_2$  выполняется условие  $t > t_{\text{кр}}$  ( $13,118 > 2,055$  и  $8,241 > 2,055$ ), а для  $a_1$  не выполняется ( $1,754 < 2,055$ ). Поэтому параметры  $a_0$  и  $a_2$  можно считать достоверно отличными от нуля, а параметр  $a_1$  нельзя.

Аналогичные результаты дает столбец  $p$ -значение для гипотезы о равенстве параметра нулю. Поскольку для  $a_0$  и  $a_2$  эта вероятность значительно меньше уровня значимости 0,05 ( $1,039 \cdot 10^{-12} < 0,05$  и  $1,362 \cdot 10^{-8} < 0,05$ ), то нулевая гипотеза отклоняется. Для  $a_1$  вероятность принятия нулевой гипотезы получилась немного больше, чем уровень значимости ( $0,09 > 0,05$ ), что не дает права отвергнуть ее. Таким образом, фактор содержания  $\text{CO}_2$  нуждается в дополнительном исследовании. Возможно, его влияние на заболеваемость носит нелинейный характер. Возможно также, что фактических данных просто было недостаточно для доказательства его влияния.

Точность регрессионной модели оценивается на основании коэффициента детерминации:  $R^2 = 0,7916$  (соответствующая строка в таблице *Регрессионная статистика*). Поскольку это значение близко к 0,8, можно говорить о том, что точность модели удовлетворительная.

Следует также оценить достоверность полученного коэффициента детерминации на основании критерия Фишера. В таблице *Дисперсионный анализ* в столбце F показано выборочное значение  $F$ -статистики, рассчитанное по формуле (3.14). Критическое значение ( $F_{\text{кр}}$ ) для уровня значимости 0,05 и степеней свободы  $k=2$  и  $n-k-1=28-2-1=25$  рассчитаем с помощью функции Excel =ФРАСПОБР(0,05;2;25).

Полученное значение равно 3,39. Поскольку выборочное значение  $F$ -статистики больше критического, нулевая гипотеза  $H_0: R^2 = 0$  отвергается. Тот же результат можно получить на основании  $p$ -значения, которое показано в столбце *Значимость F*. Это значение равно  $3,05 \cdot 10^{-9}$ , т. е. меньше уровня значимости, что также говорит о необходимости отвергнуть нулевую гипотезу и признать значимость уравнения регрессии.

Рассчитаем теперь прогноз заболеваемости, подставив в уравнение регрессии заданные в условии значения  $X_1$  и  $X_2$ . Для этого занесем эти значения в ячейки Excel (например G4 и H4, как показано на рисунке 3.24). В ячейку I4 запишем формулу уравнения регрессии, причем в качестве параметров можно использовать ссылки на соответствующие ячейки выходного диапазона. Результат расчета по этой формуле (прогнозное значение заболеваемости) составляет приблизительно 1 162.

	G	H	I
1			
2	Прогноз		
3	X1	X2	Y
4	0,7	1,5	=F\$17+F\$18*G4+F\$19*H4

Рисунок 3.24 – Расчет прогноза заболеваемости

Следует отметить, что если исключить из уравнения линейной регрессии фактор  $X_1$  и построить однофакторную модель, то получим уравнение  $\tilde{y} = 708,42 + 311,82X_2$ , причем оценки обоих параметров этой модели будут значимыми. Однако коэффициент детерминации данной модели станет несколько ниже ( $R^2 = 0,766$ ).

### Задания для самостоятельной работы

1. На основе данных, приведенных в таблице 3.26, постройте уравнение линейной регрессии и оцените его качество.
2. Используя данные таблицы 3.27, постройте уравнение линейной регрессии и оцените его качество.\*

\* Это же задание выполняется по данным таблиц 3.28–3.33.



## **СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ**

- Афифи, А.** Статистический анализ: подход с использованием ЭВМ : [пер. с англ.] / А. Афифи, С. Эйзен. – М. : Мир, 1982. – 488 с.
- Бородич, С. А.** Эконометрика : учеб. пособие / С. А. Бородич. – Минск : Новое знание, 2001. – 408 с.
- Булдык, Г. М.** Теория вероятностей и математическая статистика : учеб. пособие / Г. М. Булдык. – Минск : Выш. шк., 1989. – 285 с.
- Вентцель, Е. С.** Теория вероятностей : учеб. пособие / Е. С. Вентцель. – 7-е изд., стер. – М. : Высш. шк., 2001. – 575 с.
- Гельман, В. Я.** Решение математических задач средствами Excel : практикум / В. Я. Гельман. – СПб. : Питер, 2003. – 237 с.
- Гмурман, В. Е.** Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высш. шк., 2003. – 479 с.
- Жевняк, Р. В.** Теория вероятностей и математическая статистика : учеб. пособие / Р. В. Жевняк, А. А. Карпук, В. Т. Унукович. – Минск : Харвест, 2000. – 384 с.
- Ильина, О. П.** Статистический анализ и прогнозирование экономической информации в электронной таблице Excel 5.0. Microsoft : учеб. пособие / О. П. Ильина, Н. В. Макарова. – СПб. : СПбГУЭФ, 1996. – 140 с.
- Калинина, В. Н.** Математическая статистика : учеб. пособие / В. Н. Калинина, В. Ф. Панкин. – М. : Высш. шк., 1998. – 336 с.
- Колемаев, В. А.** Теория вероятностей и математическая статистика : учеб. пособие / В. А. Колемаев, О. В. Староверов, В. Б. Турундаевский ; под ред. В. А. Колемаева. – М. : Высш. шк., 1991. – 400 с.
- Красс, М. С.** Математика для экономических специальностей : учеб. / М. С. Красс. – 3-е изд., перераб. и доп. – М. : Дело, 2002. – 704 с.
- Лихолетов, И. И.** Высшая математика, теория вероятностей и математическая статистика / И. И. Лихолетов. – Минск : Выш. шк., 1976. – 720 с.
- Салманов, О. Н.** Математическая экономика с применением Mathcad и Excel / О. Н. Салманов. – СПб. : БХВ-Петербург, 2003. – 464 с.
- Тюрин, Ю. Н.** Статистический анализ данных на компьютере : учеб. пособие / Ю. Н. Тюрин, А. А. Макаров ; под ред. В. Э. Фигурнова. – М. : ИНФРА-М, 1998. – 528 с.
- Экономико-математические** методы и прикладные модели : учеб. пособие для вузов / В. В. Федосеев [и др.] ; под ред. В. В. Федосеева. – М. : ЮНИТИ, 2001. – 391 с.

## **СОДЕРЖАНИЕ**

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА .....	3
ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ, МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО РЕШЕНИЮ ПРИМЕРОВ, ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ....	5
1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ .....	5
1.1. Случайные события и вероятность.....	5
1.2. Способы описания и характеристики случайных величин.....	10
1.3. Основные виды распределений случайных величин .....	19
1.3.1. Биномиальное распределение .....	19
1.3.2. Распределение Пуассона .....	20
1.3.3. Равномерное распределение .....	21
1.3.4. Нормальное распределение.....	23
1.3.5. Экспоненциальное (показательное) распределение.....	28
1.3.6. Распределение $\chi^2$ (хи-квадрат).....	30
1.3.7. Распределение Стьюдента.....	31
1.3.8. Распределение Фишера .....	33
2. ВЫБОРКА И ЕЕ АНАЛИЗ .....	36
2.1. Построение и визуализация вариационного ряда.....	36
2.2. Точечные и интервальные оценки характеристик случайной величины .....	50
2.3. Общие сведения о проверке статистических гипотез .....	59
2.4. Оценка соответствия выборочных данных теоретическому закону распределения .....	66
3. АНАЛИЗ НЕСКОЛЬКИХ ВЫБОРОК .....	80
3.1. Выявление достоверности различий между двумя выборками .....	80
3.2. Дисперсионный анализ.....	100
3.3. Ковариация и корреляция.....	110
3.4. Регрессионный анализ .....	121
СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ.....	130

Учебное издание

**МЕТОДЫ  
СТАТИСТИЧЕСКОЙ ОБРАБОТКИ  
ИНФОРМАЦИИ В MS EXCEL**

**Пособие  
для студентов специальности 1-26 03 01  
«Управление информационными ресурсами»**

Авторы-составители:  
**Еськова** Оксана Ивановна  
**Авдашкова** Людмила Павловна

Редактор Е. В. Седро  
Технический редактор И. А. Козлова  
Компьютерная верстка Л. Г. Макарова

Подписано в печать 17.04.12. Бумага типографская № 1.  
Формат 60 × 84 <sup>1</sup>/<sub>16</sub>. Гарнитура Таймс. Ризография.  
Усл. печ. л. 7,67. Уч.-изд. л. 7,80. Тираж 150 экз.

Заказ №

Учреждение образования  
«Белорусский торгово-экономический  
университет потребительской кооперации».  
246029, г. Гомель, просп. Октября, 50.  
ЛИ № 02330/0494302 от 04.03.2009 г.

Отпечатано в учреждении образования  
«Белорусский торгово-экономический  
университет потребительской кооперации».  
246029, г. Гомель, просп. Октября, 50.